TREBALL DE FI DE GRAU

Grau en Matemàtiques

Títol: Modelització del nombre de sinistres d'una companyia asseguradora

Autor: Nuria Botella Garcia

Directora: Marta Pérez Cassany

Departament: Estadística i Investigació Operativa

Convocatòria: 08/01/2018

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Universitat Politècnica de Catalunya
Facultad de Matemàtiques i Estadística

Treball de Fi de Grau

Modelització del nombre de sinistres d'una companyia asseguradora

Nuria Botella Garcia

Directora: Marta Pérez Cassany

Estadística i Investigació Operativa

Índex

Introducció

El Grau en Matemàtiques proporciona una gran base teòrica d'aquesta ciència que ens permet abordar problemes d'àmbits molt diferents. Ara bé, des del meu punt de vista personal i subjectiu, també té certes mancances, com per exemple no treballar tant com a mi m'agradaria les seves aplicacions a casos reals. Per això volia realitzar un TFG que contribuís en la meva formació acadèmica des d'un punt de vista més aplicat i em permetés endinsar-me en un problema real. Aprofitant que l'empresa on vaig realitzar les pràctiques (Catalana Occident), i en la que actualment estic contractada, em proporcionava dades reals pel treball, vaig decidir endinsar-me en el món de les assegurances de cotxes, un mercat molt competitiu actualment.

Com que al Grau de Matemàtiques només es cursa una assignatura d'Estadística, una part molt important d'aquest treball ha consistit en ampliar la meva formació en aquesta àrea. Concretament, en el camp de la *modelització estadística*, que consisteix en una sèrie d'eines que permeten explicar una variable aleatòria mitjançant un conjunt de variables independents sovint controlades per l'experimentador. De fet, aquest treball de grau està estructurat en dues parts ben diferenciades. La primera part, que engloba els tres primers capítols, conté un breu resum de la teoria dels Models Lineals, els Models Lineals Generalitzats i els models importants per a les dades de comptatge. El Capítol 3 també tracta dos problemes lligats a les dades de comptatge que són la *sobredispersió* i l'*excés de zeros*.

La segona part compren els dos darrers capítols i són els capítols en els que es treballa amb dades reals. Primerament, examino la base de dades original i faig una anàlisi exploratòria per a detectar observacions anòmales i per a fixar la població objecte d'estudi. Aquesta part està exposada en el Capítol 4. Un cop fixada la variable resposta a analitzar, aplico les tècniques de modelització exposades a la primera part. Per a aquesta segona part, he consultat alguns articles de recerca que analitzen dades semblants com per exemple [22] i [23].

Per realitzar les anàlisi presentades en el Capítol 5 no només m'he hagut d'instruir en la teoria si no també en el software estadístic R que és el que hem utilitzat en el treball. A l'assignatura d'estadística vam poder treballar amb ell en les 6 pràctiques del curs, però ens vam quedar amb una visió força bàsica i, per tant, aquest TFG ha contribuït també en la meva formació en el domini d'aquest software. El treball acaba amb l'exposició de les conclusions més importants a les que hem arribat amb les nostres anàlisi.

Per a no extendre'm amb longitud del treball principal s'han inclòs tres apèndixs que contenen material que es considera important per a la bona comprensió del treball però que en certa forma és complementària.

Aquest TFG s'ha realitzat durant els 4 mesos que van de setembre a desembre compaginant-ho amb una feina a temps complert a l'esmentada Catalana Occident. Som conscients que la modelització estadística

és una cosa dinàmica i, per tant, el que aquí s'ha realitzat és una primera anàlisi de les dades que poden ser punt de partida per a anàlisis posteriors. Val a dir, que una de les coses més importants que he pogut constatar al realitzar aquest treball, ha sigut veure que la realitat és sovint molt més complexa del que descriuen els models teòrics.

Objectius

Aquest treball té objectius que es poden dividir en dos tipus: teòrics i pràctics. Els objectius teòrics pretenen l'assimilació de conceptes que no s'han vist durant els estudis de Grau. Els conceptes són de modelització estadística. Com a objectius pràctics hi figura treballar amb una base de dades real. Això vol dir portar a terme una anàlisi exploratòria de la mateixa, i un cop definida la població d'estudi, aplicar tècniques de modelització estadística per tal de treure'n informació. **Objectius teòrics:**

1. Conèixer amb més profunditat i de forma més general els Models Lineals, als quals a l'assignatura d'Estadística s'hi dedicaren no més d'un parell de setmanes.

2. Conèixer la teoria dels Models Lineals Generalitzats, que és necessària quan la variable resposta no és Normal, que és el cas d'aquest treball.

3. Conèixer i saber com abordar dos problemes molt habituals en dades discretes que són: la sobredispersió i la zero inflació.

4. Conèixer la part de software Estadístic R que implementa la teoria estudiada i que permetrà analitzar les dades.

Objectius pràctics:

1. Portar a terme una anàlisi exploratòria de la Base de Dades.

2. Analitzar la variable resposta discreta: nombre total de sinistres RC en funció de les explicatives que conté la base per tal de decidir quines d'aquestes la influencien significativament. Aquest anàlisi es realitzarà assumint diferents distribucions de probabilitat per la variable resposta.

3. Treure conclusions dels anàlisi realitzats.

Capítol 1

Models Lineals

Històricament, hi ha hagut la necessitat de plasmar mitjançant un llenguatge científic, que fos precís i clar, el comportament de fenòmens aleatoris que ens envolten amb la finalitat de poder conèixer què els afecta i, així, poder explicar-los i efectuar prediccions. Aquest llenguatge concís ha sigut el llenguatge de les matemàtiques per diverses raons: primerament és un llenguatge rigorós, ben definit i fàcil de manipular, que basa els seus procediments en resultats demostrats fa centenars d'anys i, actualment és molt assequible de programar i d'obtenir resultats numèrics mitjançant, per exemple, simulacions gràcies als ordinadors. Aquesta representació de la realitat mitjançant fórmules i conceptes matemàtics és el que coneixem com a *model matemàtic* i, si conté una component aleatòria com a *model estadístic*. Per tal de comparar diferents models per a una mateixa situació, cal tenir en compte l'anomenat *Principi de Parsimonia* que consisteix en utilitzar el menor número de variables explicatives o factors però que expliquin en gran mesura la variable, objecte d'estudi. Les variables i factors que tinguin una menor contribució seran englobades al terme *error*. Per aquest motiu, hi ha una diferència entre allò que s'observa i allò que s'espera observar segons el model.
D'aquesta manera, es pot dir que l'estructura bàsica d'un model es resumeix en la següent equació:

$$\text{"observació"} = \text{"model"} + \text{"error aleatori"}$$

Els models es realitzen amb dos clars objectius: explicar la variable resposta amb les variables explicatives més importants que la influencien i predir el seu comportament a partir d'aquesta selecció. Segons quina de les dues sigui la finalitat principal del model, afegirem més o menys factors, depenent de quina precisió volem aconseguir en la nostra predicció. En el segon cas necessitarem més variables per obtenir prediccions més exactes.
Per familiaritzar-nos amb com s'analitzen els models, abordarem en aquest capítol el cas més clàssic i alhora més senzill: el Model Lineal, que ha estat la base per a definir models posteriors. Podem consultar a la bibliografia tots els llibres o apunts que hem utilitzat per desenvolupar aquest capítol, però volem destacar especialment els apunts del professor Francesc Carmona de la Universitat de Barcelona ([5], [6], [7] i [8]) útils per entendre la teoria i amb la facilitat de l'idioma i els llibre de Models Lineals amb R de Julian J. Faraway ([3] i [4]) molt útil per familiaritzar-nos amb el software estadístic R. Una altra eina molt útil per la informació tan breu i concisa que presentava, van ser les transparències que ens va proporcionar la professora Marta Pérez Cassany [9] i que utilitza en el màster d'estadística que imparteix aquesta universitat juntament amb la UB.

1.1 Definició del Model Lineal

El *Model Lineal* (Linear Model, ML) és un model que consisteix en expressar una variable aleatòria Y (anomenada *variable depenent o de predicció*) com a combinació lineal d'unes variables de control $X_1, ..., X_m$ (anomenades *variables independents o predictores*). La relació entre ambdós tipus de variables es pot expressar com una equació de la manera següent:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \epsilon$$

Hem de destacar que el model és lineal respecte les variables explicatives, d'aquí el seu nom. Anomenant ϵ_i a la component i-èssima del vector dels errors, han de verificar les següents hipòtesi:

1. $\mathbb{E}(\epsilon_i) = 0 \quad i = 1, \ldots, n$
   En la notació matricial: $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$
2. Homocedasticitat: $Var(\epsilon_i) = \sigma^2 \quad i = 1, \ldots, n$
3. Desviacions mútuament incorrelacionades: $\mathbb{E}(\epsilon_i \cdot \epsilon_j) = 0 \quad \forall i \neq j$
   En forma matricial la condició anterior i aquesta es resumeixen en: $Var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$
4. $\epsilon_i \sim N(0, \sigma)$ i $\epsilon_1, \ldots, \epsilon_n$ independents.

Suposem que tenim n observacions independents de la variable Y que denotarem com: $y_1, ..., y_n$ i que l'observació y_i s'ha obtingut sota les condicions experimentals:

$$X_1 = x_{i1}, \quad X_2 = x_{i2}, \quad \ldots \quad X_m = x_{im}$$

declarades prèviament en el moment de dissenyar l'experiment. Aleshores l'equació del model s'escriu:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \epsilon_i \quad i = 1, \ldots, n$$

i, en forma matricial:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

que en forma comprimida és igual a:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.1}$$

Cal remarcar que els elements de la matriu $\mathbf{X}$ són coneguts ja que, com s'ha dit amb anterioritat, han estat fixats al dissenyar l'experiment. D'aquí que la matriu $\boldsymbol{X}$ s'anomeni *matriu de disseny*.
Tenint en compte les hipòtesi del model lineal sobre els errors, l'equació (1.1) és equivalent a:

$$\boldsymbol{Y}|\boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) \tag{1.2}$$

Als models que verifiquen aquesta condició se'ls anomena models lineals.
Com que un dels nostres objectius és predir la variable resposta $\boldsymbol{Y}$, fins i tot sota condicions diferents de les experimentals, hauríem de tenir determinats completament els coeficients que apareixen en l'equació

(1.2). Malauradament desconeixem el valor del vector de paràmetres $\boldsymbol{\beta}$ i la variància σ^2 i, per tant, caldrà estimar-los a partir de les observacions. En la següent secció ens centrarem en com podem trobar el seus valors.

1.2 Estimació dels paràmetres del model

Clarament, no podem calcular els valors reals dels coeficients desconeguts perquè els models són aproximats, així que l'alternativa és estimar-los. A més, cal recordar que el que s'observa no és determinista i, per tant, en les mateixes condicions experimentals es poden obtenir valors diferents de la variable dependent. En inferència estadística, *estimar un coeficient* consisteix en aplicar un conjunt de tècniques amb la finalitat de donar un valor aproximat a un paràmetre del model a partir d'una mostra. Hi ha tres tipus d'estimació: estimació puntual, estimació per intervals i l'estimació bayesiana. El valor aproximat rep el nom d'*estimador* i el denotem amb un barret, per exemple, indicaríem l'estimador del vector de paràmetres $\boldsymbol{\beta}$ com $\hat{\boldsymbol{\beta}}$. En aquesta secció aplicarem les dues primeres tècniques d'estimació per calcular les estimacions dels paràmetres del model lineal.

1.2.1 Estimació puntual

L'*estimació puntual* és una tècnica d'estimació, que consisteix en aproximar el valor desconegut mitjançant un únic valor, obtingut a partir d'una formula determinada. Les tècniques utilitzades habitualment són: el mètode dels moments, el criteri dels mínims quadrats i el mètode de la màxima versemblança.
En els següents apartats aplicarem els dos darrers mètodes en el cas del model lineal. I tal com es veurà ambdós donen lloc a la mateixa estimació.

1.2.1.1 Criteri dels Mínims Quadrats

El *Criteri dels Mínims Quadrats* (Minim Least Squares, MLS) és un mètode que adapta el model estadístic a un conjunt de dades en els casos en que el valor predit pel model s'expressa linealment en termes dels paràmetres desconeguts. Utilitza l'algoritme dels *mínims quadrats*, tècnica en que la millor aproximació és aquella que minimitza la suma del quadrat de les diferències entre les dades i els seus valors predits pel model.
És a dir, consisteix en minimitzar la suma del quadrat dels errors:

$$SSE = \epsilon^T \epsilon = \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \ldots - \beta_m x_{im})^2$$

escrit en forma matricial:

$$\epsilon^T \epsilon = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{1.3}$$

Desenvolupant el producte de l'equació (1.3) obtenim:

$$e^T e = \boldsymbol{Y}^T \boldsymbol{Y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{Y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{\beta}$$

Per trobar la $\boldsymbol{\beta}$ que minimitza l'equació, hem de derivar-la respecte de $\boldsymbol{\beta}$ i igualar-la a zero, obtenint:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left(\boldsymbol{Y}^T \boldsymbol{Y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{Y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{\beta} \right) = -2\boldsymbol{X}^T \boldsymbol{Y} + 2\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} = 0$$

D'aquí que l'estimador haurà de verificar l'equació:

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y} \tag{1.4}$$

L'equació (1.4) rep el nom d'*equació normal* i aïllant $\hat{\boldsymbol{\beta}}$ tenim el seu valor:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{1.5}$$

Observem que aquesta expressió només té sentit quan el producte de matrius $\boldsymbol{X}^T\boldsymbol{X}$ és invertible, i això succeeix quan el rang de $\boldsymbol{X}$ és màxim. Això sí, quan no estem en aquest cas hi ha una solució per poder calcular l'expressió de l'estimador, calcular la g-inversa de $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}$, és a dir, la que verifiqui la condició $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$; on $\boldsymbol{A}^-$ denota la g-inversa.
L'inconvenient d'haver de calcular la g-inversa és que aleshores hi han infinites possibles solucions per a l'estimador $\hat{\boldsymbol{\beta}}$, però l'avantatge és que el vector predit pel model, denotat per $\hat{\boldsymbol{Y}}$ i que es calcula com:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

és igual per a qualsevol dels $\hat{\boldsymbol{\beta}}$ usats. Utilitzar una g-inversa o una altra només portarà a una diferent interpretació dels paràmetres del model.
Un cop calculat l'estimador de $\boldsymbol{\beta}$, només ens queda estimar el coeficient de la variància, σ^2.
L'equació (1.2) particularitzada a una observació, i entenent que $x_{0j} = 1 \forall j$, és:

$$y_i \sim N\left(\sum_{i=0}^{p-1} x_{ij}\beta_j, \sigma^2\right) \tag{1.6}$$

Si la estandaritzem, és a dir, si li restem el seu valor esperat i dividim per la seva Deviança estàndard s'obté:

$$\frac{y_i - \hat{y}_i}{\sigma} \sim N(0,1)$$

La suma de normals estàndards al quadrat segueix, per definició, una distribució chi-quadrat. Així doncs,

$$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sigma^2} = \frac{\sum_{i=1}^n (y_i - (x\hat{\beta})_i)^2}{\sigma^2} \sim \chi^2_{n-p}$$

L'esperança d'una χ^2 és igual al seu nombre de graus de llibertat. Així doncs, igualant l'esperança al valor que havíem observat s'obté:

$$\frac{\sum_{i=1}^n (y_i - (x\hat{\beta})_i)^2}{\sigma^2} = n - p$$

d'on s'obté que:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - (x\hat{\beta})_i)^2}{n-p} \tag{1.7}$$

Aquest estimador de la variància es denota per s^2.

1.2.1.2 El mètode de la màxima versemblança

El *Mètode de la Màxima Versemblança* (Maximum Likelihood Estimation, MLE) és una tècnica que s'utilitza per estimar els paràmetres d'un model, es fa trobant els valors que maximitzen la funció de versemblança a partir de les observacions. La principal diferència amb el mètode anterior, pel que respecta a l'estructura dels paràmetres $\boldsymbol{\beta}$ és que es necessita especificar la distribució que segueixen les observacions de la variable resposta, degut a que la funció de versemblança es calcula a partir de la funció de densitat conjunta.
A l'Apèndix A.1 queda explicat el procediment general del MLE i ara el particularitzarem al cas en que les observacions y_i siguin normals (és a dir, que segueixen l'equació (1.6)), que és el cas del ML.
La funció de densitat d'y_i és:

$$f_{y_i;\beta,\sigma^2} = \frac{1}{\sqrt{\sigma}\sqrt{2\pi}} e^{-\frac{(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j)^2}{2\sigma^2}}$$

La funció de versemblança es defineix com:

$$L(\beta, \sigma^2; y) = \prod_{i=1}^{n} f_{y_i;\beta,\sigma^2} = \frac{1}{(\sigma^2)^{n/2}(\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{i=0}^{p-1} x_{ij}\beta_j\right)^2} \tag{1.8}$$

Atès que el màxim d'una funció coincideix amb el màxim del seu logaritme, prenent logaritmes a (1.8) s'obté l'anomenada funció de versemblança que és igual a:

$$l(\beta) = -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{i=0}^{p-1} x_{ij}\beta_j\right)^2 \tag{1.9}$$

El vector de paràmetres $\boldsymbol{\beta}$ que maximitza a la log-versemblança s'obtindrà resolent el sistema resultant d'igualar les derivades parcials a zero. Aquest sistema d'equacions s'anomena *equacions normals* i és igual a:

$$\frac{\partial}{\partial\beta_j} l(\beta) = \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i - \sum_{i=0}^{p-1} x_{ij}\beta_j)x_{ij} = 0 \quad j = 1 \div p-1$$

Si ho escrivim en forma matricial és més fàcil veure quina relació té amb el criteri MLS:

$$\frac{1}{\sigma^2}(\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})) = 0 \Rightarrow \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = 0 \Rightarrow \boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \Rightarrow \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

i, per tant l'estimador de $\boldsymbol{\beta}$ és idèntic per als dos mètodes.
Per finalitzar la secció, calcularem l'estimador de σ^2 amb aquest mètode. Escrivint la funció de log-versemblant en funció de σ^2 tenim que de (1.9):

$$l(\sigma^2) = -\frac{n}{2}\log\sigma^2 + \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{i=0}^{p-1} x_{ij}\beta_j\right)$$

Derivant respecte de σ^2 i igualant la derivada a zero s'obté que l'estimador màxim versemblant de σ^2 és:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{i=0}^{p-1} x_{ij}\beta_j\right) = \left(1 - \frac{p}{n}\right)s^2$$

observis que pel que respecte a σ^2, els dos estimadors poden diferir sobretot si n no és gaire gran. Per n gran, els dos estimadors són pràcticament el mateix.

Hem vist que amb els dos mètodes d'estimació puntual aconseguim la mateixa expressió de l'estimador del vector de paràmetres $\boldsymbol{\beta}$, però d'altra banda, ambdues variàncies només coincideixen assimptòticament (és a dir, per a n grans).
El teorema de Guass-Markov ens dóna la resposta a la pregunta: quins dels dos mètodes és el millor? Al Apèndix A.1 podrem trobar el teorema i els conceptes necessaris per enunciar-lo. Bàsicament, aquest teorema ens està dient que l'estimador amb millors propietats és l'obtingut mitjançant MLE. Però, quan la distribució de les observacions és normal, aleshores ambdós estimadors coincideixen, que és el que succeeix en el nostre cas. En el cas de l'estimació de la variància, té millors propietats l'obtingut mitjançant MLS per ser no esbiaixat, és a dir, $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$. Per aquest motiu sempre es pren (1.7) com a estimador de la variància dels errors.

1.2.1.3 Hipòtesi sobre els paràmetres

A la introducció, hem comentat que una variable resposta Y pot estar influenciada per nombroses variables explicatives, però degut a que ens interessa aconseguir un model senzill, només volem considerar aquelles que tinguin una contribució significativa. Per aquest motiu, la resta seran agrupades en el terme que s'anomena *error*.
Aquest fet fa especialment important estudiar la significació dels paràmetres del model, atès que si algun d'ells no és significativament diferent de zero, voldria dir que la variable explicativa associada pot eliminar-se del model per no tenir una influència important en la resposta.
Avaluar la significació d'un paràmetre es realitza a través del següent test d'hipòtesi:

$$\begin{cases} H_0: & \beta_i = 0 \\ H_1: & \beta_i \neq 0 \end{cases}$$

i el test s'efectua $\forall i \in 0, 1, \ldots, p-1$. La hipòtesi nul·la es rebutja quan:

$$\left| \frac{\hat{\beta}_i}{\sqrt{(X^T X)^{-1}_{ii}}} \right| \geq t_{\alpha/2, n-p}$$

essent $t_{\alpha/2,n-p}$ el valor que deixa una probabilitat per sobre igual a $\alpha/2$ en la distribució t-d'Student amb $n-p$ graus de llibertat. Important remarcar que α s'anomena *nivell de significació del test* i correspon a la probabilitat de rebutjar la hipòtesi nul·la quan és certa. És a dir:

$$\alpha = Pr(\text{rebutjar } H_0 | H_0 \text{ certa})$$

Quan H_0 no sigui rebutjada, podrem eliminar la variable X_i del model.

1.2.1.4 Inferència sobre els paràmetres estimats

Si fem memòria, el nostre objectiu principal era estimar la variable resposta Y, però teníem el petit inconvenient de que l'equació que la definia (1.2) no estava completament determinada perquè hi havia coeficients que desconeixíem. Arribats a aquest punt, ja coneixem tots els factors i som capaços d'estimar Y:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y} \tag{1.10}$$

on $\boldsymbol{H}$ s'anomena *hat matrix*, perquè és la matriu per la qual hem de multiplicar la Y per obtenir $\hat{Y}$.
L'error aleatori és:

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$$

i la suma del quadrat dels errors, o suma de quadrats residuals, serà igual a:

$$SSE = \boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{Y}^T\boldsymbol{Y} - 2\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y}$$

A l'Apèndix A.1 s'explica que tant els paràmetres estimats $\hat{\boldsymbol{\beta}}$, com els valors predits, $\hat{\boldsymbol{Y}}$, són vectors aleatoris amb distribució Normal, i se'n determinen l'esperança i la variància.
L'estimació puntual no sempre és la més útil, de fet, a la practica és moltes vegades preferible conèixer un interval que contingui amb probabilitat alta el valor del paràmetre que la pròpia estimació puntual. Per exemple, en el cas d'una asseguradora, donat un perfil de conductor i vehicle, és més interessant conèixer el rang del cost d'un sinistre que pugui tenir, que el valor propi del cost del sinistre. Per aquest motiu, en la següent secció estudiarem com calcular intervals de confiança.

1.2.2 Intervals de confiança (IC)

L'*estimació per intervals* consisteix en obtenir els límits entre els quals es troba el veritable valor del paràmetre amb una probabilitat donada. Aquest interval s'anomena *interval de confiança*. Es denota com $1-\alpha$ el *nivell de confiança de l'interval*, que consisteix en la mesura de la probabilitat de que el paràmetre estigui a l'interval calculat. Com hem dit, aquesta mesura l'estableix l'estadístic abans de realitzar els càlculs i normalment es considera igual a 95% o 90%.
Realitzarem els càlculs pel cas de l'estimació d'una component β_i del vector de paràmetres $\boldsymbol{\beta}$. Hem vist, a l'Apèndix A.1 que aquest estimador segueix una distribució normal multivariant (A.1).
Si la tipifiquem tenim:

$$\frac{\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} - \boldsymbol{\beta}_{\boldsymbol{i}}}{\sigma\sqrt{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{ii}}} \sim N(0,1)$$

Atès que la variància és desconeguda, l'estimem a partir de la suma de quadrats residual tal com figura a (1.7). A l'estimar la variància, la distribució de l'estadístic anterior passa de ser una Normal a ser una t-d'Student. Així doncs,

$$\frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{s\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}}} \sim t_{n-p}$$

Per calcular l'IC de $\boldsymbol{\beta}$ només hem d'imposar la seva definició, i aïllar β_i. Si $t_{n-p,\alpha/2}$ és el valor que deixa una probabilitat per sobre igual a $\alpha/2$ en una distribució t-d'Student amb $n-p$ graus de llibertat llavors es té que:

$$Pr\left(-t_{n-p,\frac{\alpha}{2}} < \frac{\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i}{s\sqrt{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{ii}}} < t_{n-p,\frac{\alpha}{2}}\right) = 1-\alpha$$

el que equival a:

$$Pr\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} - t_{n-p,\frac{\alpha}{2}}s\sqrt{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{ii}} < \boldsymbol{\beta}_{\boldsymbol{i}} < \hat{\boldsymbol{\beta}}_{\boldsymbol{i}} + t_{n-p,\frac{\alpha}{2}}s\sqrt{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{ii}}\right) = 1-\alpha$$

Finalment, l'IC del paràmetre $\boldsymbol{\beta}_i$ és:

$$\left[\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} - t_{n-p,\frac{\alpha}{2}}s\sqrt{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{ii}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{i}} + t_{n-p,\frac{\alpha}{2}}s\sqrt{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{ii}}\right]$$

Important observar que el test d'hipòtesi esmentat en la subsecció 1.3.1.3 també es pot fer a partir de l'IC per a β_i, rebutjant H_0 quan l'interval no conté el zero.

1.3 Bondat d'ajust del model

Un cop ja tenim definit el nostre model i amb tots els paràmetres estimats, hem de comprovar la seva *bondat d'ajust*, procés que consisteix en comprovar que el model resultant verifica les hipòtesi del ML. També mesura lo bé que s'ajusta al conjunt d'observacions.
A continuació, mostrarem les eines principals que s'utilitzen per comprovar les hipòtesi del ML:

- **Linealitat:** per veure que podem aproximar les nostres observacions mitjançant una recta, s'ha de verificar que la variable resposta i les variables explicatives estan relacionades linealment. Per comprovar-ho, es fa servir un scatter-plot o núvol de punts de Y vs X, on s'ha d'observar aquesta característica entre les variables. A l'Apèndix A.1 podem consultar quina forma hauria de tenir aquesta gràfica. Aquesta eina és molt útil, sobretot en *regressió lineal simple* que és la té només una variable explicativa.

- **Homocedasticitat:** una altra condició que es demana és la igualtat de variàncies. El més habitual, és comprovar aquesta condició gràficament a partir dels residus estimats. Es realitza un scatter-plot entre l'estimació dels residus $\hat{e}$ i l'estimador de la variable resposta $\hat{y}$ i no s'ha d'observar cap patró en la variabilitat de les observacions. L'homocedasticitat es té quan l'amplada dels residus és pràcticament la mateixa en tot el tram. A l'Apèndix A.1 podem trobar exemples de situacions que tenen la propietat d'homocedasticitat i la d'heterocedasticitat.

- **Normalitat dels residus**. De la mateixa manera que abans, el més usual és verificar-ho gràficament. En aquest cas, s'ha d'observar una relació lineal en el qq-plot dels errors estimats. En l'Apèndix A.1 podem trobar exemples de qq-plots on els residus són normals i exemples en que no ho són.

Per finalitzar el procés de bondat d'adust, només cal mesurar la relació entre el valors observats i els valors predits. Per fer-ho, es calcula l'indicador R^2:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

El seu rang és $0 \leq R^2 \leq 1$, els valors propers a 1 indiquen una millor aproximació. De fet, aquesta mesura de bondat d'ajust consisteix en la proporció de la variabilitat total que hi ha en la variable resposta que és explicada pel nostre treball. L'avantatge d'utilitzar aquest indicador és que no té unitats i, per tant, és fàcil d'entendre i de comparar. L'inconvenient d'aquest estadístic és que com més variables explicatives afegeixis major serà el seu valor (és a dir, més proper a 1) i, per tant, se suposa que millor ajustat està el model.
Però, com hem insistit al llarg del capítol, ens interessen models senzills que siguin capaços d'explicar la variable resposta amb el menor número de variables predictores. Per això, existeix un altre estadístic conegut com R^2 *ajustada,* R^2_{adj} que té la mateixa funció que el coeficient R^2 però, a més, penalitza quan s'afegeixen variables X_i, atès que té en compte el nombre de paràmetres del model. Un avantatge del R^2_{adj} és que ens permet comparar diferents models per a un conjunt de dades. En el sentit que donats dos

models, serà millor aquell que tingui un R^2_{adj} més gran. Aquest coeficient es calcula com:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p}$$

A l'Apèndix A.1, hem afegit un exemple del cas més senzill del ML: la regressió lineal simple. Ho hem fet, per a que quedin clars els procediments i els càlculs en un exemple.

Ara que ja hem acabat el capítol, hem fet un breu resum de la teoria dels ML i hem vist un exemple pràctic, ens hauríem d'haver adonat de la seva limitació. És inevitable que ens sorgeixin preguntes com:

- Que passa si la variable resposta del nostre problema no segueix una distribució normal?
- I si l'esperança de la variable resposta no és a funció que depèn linealment de les variables explicatives?
- I si la variància no és constant, si no que canvia en funció de les condicions experimentals?

Respondre aquestes preguntes serà l'objectiu del proper capítol.

Capítol 2

Models Lineals Generalitzats

En el capítol anterior, hem pogut veure les limitacions dels ML i, com hem dit al final, ara volem trobar una extensió que permeti aplicar mètodes anàlegs a situacions més generals. Per exemple, situacions en que la distribució de la variable resposta no sigui una normal o en que la funció esperança no sigui la que està relacionada linealment amb la matriu $\boldsymbol{X}$.
Aquestes situacions es donen amb molta freqüència, de fet en aquest treball la variable resposta és el nombre de sinistres d'una companyia asseguradora i tractant-se d'una variable discreta, no té sentit que es distribueixi segons una Normal.
Per això, en aquest capítol ens centrarem en presentar de forma resumida la teoria dels Models Lineals Generalitzats: les seves característiques, en quines distribucions s'aplica i la bondat d'ajust del model. Per aquest capítol volem destacar dos llibres de la bibliografia que han estat especialment importants: el llibre de l'Annette J. Dobson ([11], [12], [13] i [14]), molt útil com a introducció als GLM tal i com diu el títol i el llibre de McCullagh i Nelder ([15] i [16]), molt utilitzat en altres projectes per ser un dels primers en tractar aquest tema de forma tan complerta.

2.1 Definició del Model Lineal Generalitzat

Els *Models Lineals Generalitzats* (MLG) són una extensió dels ML i es caracteritzen per tenir les tres components següents:

- **La component aleatòria:** identifica la variable resposta Y_i i defineix quina distribució segueix, serà una de les que pertanyen a les anomenades famílies exponencials i la seva esperança es denotarà:

$$\mathbb{E}(Y_i) = \mu_i$$

- **La component sistemàtica o determinista:** especifica les variables explicatives del model que intervenen en la funció predictora lineal. Engloba el vector de paràmetres $\boldsymbol{\beta}$ i la matriu de les condicions experimentals $\boldsymbol{X}$. Habitualment, es denota de la manera següent:

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$$

- **La *funció link*** és aquella funció que com el seu nom indica, relaciona el valor esperat de la variable resposta i la component sistemàtica. Per tant, és la funció g tal que:

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta}$$

En les següents seccions s'expliquen amb més detall les components anteriors.

2.1.1 La família exponencial

La *família exponencial* és un conjunt format per aquelles distribucions de probabilitat en que la seva funció de densitat $f(y, \theta)$ es pot expressar de la següent manera:

$$f(y;\theta) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)} \tag{2.1}$$

on, les funcions a, b i c són conegudes i θ rep el nom de *paràmetre canònic*. Al paràmetre ϕ se'l coneix com a paràmetre de dispersió, ens ha sortit en el capítol anterior però amb la notació σ^2, ja que la Normal pot escriure's com (2.1) amb $\phi = \sigma^2$.
L'esperança i la variància d'una variable aleatòria $\boldsymbol{Y}$ amb distribució (2.1) poden trobar-se aplicant les següents formules, que les relacionen amb la primera i segona derivada de la funció de log-versemblança:

$$\begin{aligned} &\mathbb{E}\left(\tfrac{\partial l}{\partial \theta}\right) = 0 \\ &\mathbb{E}\left(\tfrac{\partial^2 l}{\partial \theta^2}\right) + \mathbb{E}\left(\tfrac{\partial l}{\partial \theta}\right)^2 = 0 \end{aligned} \tag{2.2}$$

La log-versemblança per a la distribució de probabilitat amb densitat (2.1) és igual a: És a dir, necessitem calcular la funció de log-versemblança, recordem que era el logaritme de la funció de densitat i en aquest cas segueix la següent expressió:

$$l(\theta; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

L'esperança l'obtindrem substituint la seva primera derivada en la primera igualtat de l'equació (2.2):

$$\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = \mathbb{E}\left(\frac{y - b'(\theta)}{a(\phi)}\right) = \frac{\mu - b'(\theta)}{a(\phi)} = 0 \Rightarrow \mathbb{E}(Y) = \mu = b'(\theta) \tag{2.3}$$

on, $b'(\theta)$ indica la primera derivada de b respecte θ.
El càlcul de la variància segueix el mateix procediment que acabem de veure: hem de substituir la primera i segona derivada de la funció log-versemblant en la segona expressió de (2.2):

$$\begin{gathered} \mathbb{E}\left(\tfrac{\partial^2 l}{\partial \theta^2}\right) + \mathbb{E}\left(\tfrac{\partial l}{\partial \theta}\right)^2 = \mathbb{E}\left(\tfrac{-b''(\theta)}{a(\phi)}\right) + \mathbb{E}\left(\tfrac{y - b'(\theta)}{a(\phi)}\right)^2 = \tfrac{-b''(\theta)}{a(\phi)} + \tfrac{Var(Y)}{a^2(\phi)} = 0 \\ \Downarrow \\ Var(Y) = b''(\theta)a(\phi) \end{gathered}$$

on $b''(\theta)$ indica la segona derivada de b respecte θ. Aquesta equació està formada pel producte de dues funcions: una depenent del paràmetre canònic θ i relacionada directament amb l'esperança de $\boldsymbol{Y}$ (aquesta rep el nom de *funció variància*) i l'altra que depèn del paràmetre de dispersió ϕ. Difereix amb la variància del capítol anterior en que no és constant, és a dir, deixa de ser necessària la condició d'homocedasticitat, degut a que la variància de la resposta depèn de l'esperança μ que varia segons les condicions experimentals.
La funció variància la denotarem com $Var(\mu)$, és a dir,

$$Var(\mu) = b''(\theta)$$

2.1.2 La funció link

La *funció link* és, per definició, una funció monòtona i diferenciable un a un, que permet relacionar l'esperança de la variable resposta $\boldsymbol{\mu}$ amb el predictor lineal $\boldsymbol{\eta}$, és a dir:

$$g(\boldsymbol{\mu}) = \boldsymbol{X\beta} = \boldsymbol{\eta} \tag{2.4}$$

Com que la funció g és invertible, podem aïllar $\boldsymbol{\mu}$ d'aquesta expressió:

$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\boldsymbol{X\beta}),$$

la qual cosa ens permetrà estimar la resposta esperada en funció de les condicions experimentals.
Si bé cada funció de densitat de la forma (2.1) pot utilitzar-se amb qualsevol funció link, hi ha una funció que és especialment important que és l'anomenada *link canònic*. El link canònic és aquell que expressa el paràmetre canònic $\boldsymbol{\theta}$ de la funció de densitat, en funció del predictor lineal $\boldsymbol{\eta}$, és a dir, la funció canònica de link succeeix quan $\boldsymbol{\theta} = \boldsymbol{\eta}$. De fet, utilitzar la funció link canònic té diverses avantatges, com l'existència d'un estadístic suficient, $\boldsymbol{X}^T\boldsymbol{Y}$ i la simplificació dels càlculs per obtindre els estimadors màxim versemblants de $\boldsymbol{\beta}$.
Hem de matisar que, tot i els avantatges que té la funció link canònic, en molts casos s'utilitzen altres funcions per a definir el model, com és en el cas de la distribució Binomial.
A l'apèndix A.2 podem veure que els ML són un cas particular dels MLG. Concretament, un MLG és un ML quan la distribució resposta és la Normal i la funció link és la identitat.

2.2 Estimació dels paràmetres del model

El mètode que s'utilitza en els MLG per estimar el vector de paràmetres $\boldsymbol{\beta}$ és el de màxima versemblança. Ara bé, trobar l'estimador màxim versemblant de $\boldsymbol{\beta}$ equival a fer uns *Mínims quadrats ponderats iteratius* (Iterative Weighted Least Squares, IWLS). Consisteix en un procés iteratiu on intervenen dues variables auxiliars: $\boldsymbol{z}$ considerada la variable depenent en aquest procediment i $\boldsymbol{W}$ la matriu de pesos que depèn del valor ajustat de $\boldsymbol{\mu}$, el qual denotarem com $\hat{\boldsymbol{\mu}}$.
Una iteració d'aquest procés ve representada a la Figura 2.1:

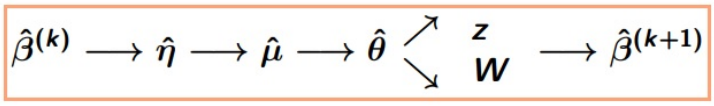

Figura 2.1: Esquema de l'algoritme del $IWLS$

Considerem que $\hat{\boldsymbol{\beta}}$ és l'estimador actual del vector de paràmetres, aleshores podem calcular el predictor lineal $\hat{\boldsymbol{\eta}}$ aplicant la definició, vista quan hem parlat de les components dels MLG:

$$\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

A partir d'aquí, podem calcular $\hat{\boldsymbol{\mu}}$ fent la inversa de la funció link:

$$\hat{\boldsymbol{\mu}} = g^{-1}(\hat{\boldsymbol{\eta}}) \tag{2.5}$$

i atès que $\boldsymbol{\mu}$ i $\boldsymbol{\theta}$ estan relacionades per (2.3), a partir de $\hat{\boldsymbol{\mu}}$ obtindrem $\hat{\boldsymbol{\theta}}$.
Usant aquests coeficients podem calcular el valor de les dues funcions auxiliars introduïdes pel mètode, la funció $\boldsymbol{z}$ quedaria:

$$\boldsymbol{z} = \hat{\boldsymbol{\eta}} + (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \frac{d\eta}{d\mu}\Big|_{\hat{\mu}}$$

on, $\frac{d\eta}{d\mu}\Big|_{\hat{\mu}}$ indica que la derivada de $\boldsymbol{\eta}$ respecte de $\boldsymbol{\mu}$ avaluada en l'actual estimador, $\hat{\boldsymbol{\mu}}$.
La funció pes quedaria:

$$\boldsymbol{W} = \left(\frac{d\eta}{d\mu}\right) Var^{-1}\Big|_{\hat{\mu}}$$

on $Var^{-1}\Big|_{\hat{\mu}}$ indica la inversa de la funció variància avaluada en l'actual estimador de $\boldsymbol{\mu}$.
Finalment, obtenim l'estimador de $\hat{\boldsymbol{\beta}}^{(km)}$, corresponent a la iteració (km), substituint $\boldsymbol{z}$ i $\boldsymbol{W}$ a la següent equació:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}$$

on $\boldsymbol{X}$ és la matriu de les condicions experimentals.
Atès que volem obtenir la menor variabilitat possible, aquest procediment es repetix fins que les diferències entre dos vectors $\boldsymbol{\beta}$ consecutius sigui prou petita.
Un cop obtingut l'estimador del vector $\boldsymbol{\beta}$, podem estimar la variable resposta $\hat{\boldsymbol{Y}}$:

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{X}\hat{\boldsymbol{\beta}})$$

Hem de fer dues observacions sobre aquest algoritme. A la primera iteració no tenim cap estimació de $\boldsymbol{\beta}$, així que en la bibliografia es recomana començar en el segon pas amb una estimació de $\boldsymbol{\mu}$ feta a partir de les observacions. En segon lloc, calcular inverses té un cost molt car i ralentitza el procediment, per això, si disposem de la funció link canònic podem passar directament al tercer pas i evitar aquest càlcul definit a l'equació (2.5).
Finalitzada aquesta secció, ens trobem que ja hem fitat el model, però de la mateixa manera que en el capítol anterior, el següent pas és comprovar que el model s'ajusta adequadament a les nostres observacions i això és el que veurem en la següent secció.

2.3 Bondat d'ajust del model

Els coeficients de bondat d'ajust tenen per objectiu mesurar la discrepància existent entre el vector de valors observats i el de predits o estimats. La intenció és aconseguir que aquesta discrepància sigui la més petita possible amb un nombre no gaire elevat de paràmetres.
Hi han diversos estadístics per calcular-la, en aquest capítol veurem els tres més importants: la Deviança, l'estadístic generalitzat de Pearson i l'AIC.

2.3.1 La Deviança

D'entre tots els models possibles n'hi ha dos de destacats: el *model nul* que és aquell que considera una resposta constant per a totes les variables $y's$ i, per tant, té un únic paràmetre, i el *model complet* que és aquell que si tens n observacions aleshores considera n paràmetres per explicar-lo. El primer és massa simple i el segon és poc informatiu, com hem repetit en diverses ocasions al capítol anterior, els models

s'enfoquen en explicar satisfactòriament la variable resposta amb el mínim de variables explicatives possibles, com diu el Principi de Parsimonia.
La *Deviança* (deviance) és un estadístic que s'utilitza per a mesurar la bondat d'ajust d'un model comparant-lo amb la bondat d'ajust obtinguda amb el model complet que es pot considerar perfecte. El test d'hipòtesi que estaríem fent en aquesta ocasió seria:

$$\begin{cases} H_0: & \text{el nostre model} \\ H_1: & \text{el model complert} \end{cases}$$

El càlcul d'aquest estadístic relaciona la funció de log-versemblança del model complert i la del model de p-paràmetres, però abans necessitem establir quina notació utilitzarem en cada cas. En la següent taula hem escrit la notació que utilitzarem per a cada model:

	Model de p paràmetres	**Model Complert**
Funció log-versemblant	$l(\hat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})$	$l(\boldsymbol{y}, \phi; \boldsymbol{y})$
Estimador del paràmetre canònic	$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{\mu}})$	$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\boldsymbol{y})$

Taula 2.1: Notació utilitzada per a definir la Deviança

El càlcul de la *Deviança escalada* és proporcional a dos vegades la diferència de la funció log-versemblant del model complert i la del model de p paràmetres, és a dir:

$$D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = 2\big(l(\boldsymbol{y}, \phi; \boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})\big)$$

Es relaciona amb la Deviança de la següent manera:

$$D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \frac{D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})}{\phi}$$

si la hipòtesi nul·la de que el nostre model és l'apropiat és certa, $D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ haurà de provenir d'una distribució χ^2_{n-p}. Així doncs, rebutgem el nostre model quan:

$$D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \geq \chi^2_{\alpha, n-p}.$$

Una aplicació molt important d'aquest estadístic és determinar el model que millor fita les observacions i amb el menor número de variables explicatives en el cas que tinguem models niats. Si considerem dos models m_1 i m_2 amb p_1 i p_2 paràmetres respectivament, direm que m_1 *està niat a* m_2 (és a dir, $p_2 >> p_1$) si el model m_2 inclou tots els paràmetres de m_1 i alguns més. En aquest cas, s'esta realitzant el test d'hipòtesi:

$$\begin{cases} H_0: & \text{model } m_1 \text{ de } p_1 \text{ paràmetres} \\ H_1: & \text{model } m_2 \text{ de } p_2 \text{ paràmetres} \end{cases}$$

rebutjarem H_0 serà quan la diferència de les dues Deviances sigui superior a $\chi^2_{p_1-p_2}$:

$$D^*_1(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) - D^*_2(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \geq \chi^2_{p_1-p_2}$$

2.3.2 L'Estadístic generalitzat de Pearson

L'*Estadístic generalitzat de Pearson* ,$\boldsymbol{X^2}$, mesura la discrepància entre una distribució observada i una teòrica, es calcula de la següent manera:

$$\boldsymbol{X^2} = \sum \frac{(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^2}{Var(\hat{\boldsymbol{\mu}})}$$

on, $Var(\hat{\boldsymbol{\mu}})$ és l'estimació de la funció variància. En el cas d'una distribució Normal, aquesta expressió coincideix amb la suma dels errors al quadrat, perquè la funció variància de la normal és igual a la unitat. Tant la Deviança com l'estadístic de Pearson X^2 segueixen exactament una distribució χ^2_{n-p} en el cas de la distribució Normal, i per a qualsevol altra distribució podem dir que segueix una χ^2_{n-p} asimptòticament. L'avantatge d'utilitzar la Deviança com a mesura de discrepància és que permet comparar models niats, en canvi, l'estadístic de Pearson X^2 no. Ara bé, aquesta darrera és de més fàcil interpretació.

2.3.3 Criteri d'informació d'Akaike, AIC

A la pràctica, es poden considerar diversos models per a un mateix conjunt de dades, i el nostre objectiu seria determinar quin de tots ells fita millor les observacions i amb el menor número de variables explicatives. El *criteri d'informació d'Akaike*(Akaike Information Criterion, AIC) és un estimador de la qualitat relativa d'un model per a un conjunt de dades donat, és a dir, estima la qualitat d'un model tenint en compte el seu número de paràmetres. Es calcula de la següent manera:

$$AIC = 2p - 2\log L(\hat{\theta}|y) \tag{2.6}$$

El millor model serà aquell que minimitzi el valor de l'equació (2.6). L'avantatge d'utilitzar aquest estimador és que recompensa la bondat d'ajust i penalitza quan creix el número de paràmetres, tal i com feia R^2_{adj} dels ML.

2.4 Els residus

Recordem que els *residus* són la diferència entre el valor observat i el seu estimador i són útils per comprovar com s'adapta el model a les observacions, i per detectar valors anomals que s'han d'estudiar amb més profunditat.
Per tant, necessitem tenir en els MLG una generalització del terme residual i que tingui propietats semblants al residu dels ML. En aquest apartat estudiarem dos tipus de residus: residu de Pearson i el residu de la Deviança.

2.4.1 El residu de Pearson

El residu de Pearson es defineix com la diferència entre el valor observat i el valor estimat, escalat per la Deviança estàndard estimada, la seva expressió és:

$$r_P = \frac{\boldsymbol{y} - \boldsymbol{\mu}}{\sqrt{Var(\boldsymbol{\mu})}}$$

Rep aquest nom perquè en el cas d'una distribució de Poisson la suma al quadrat dels r_P's és igual a l'estadístic generalitzat de Pearson, és a dir:

$$\sum r_P^2 = X^2$$

El principal inconvenient d'utilitzar aquest residu és que en el cas de distribucions no Normals acostuma a ser esbiaixat, per això necessitem alternatives que tinguin unes característiques semblants a les dels residus dels ML.

2.4.2 El residu de la Deviança

En la secció de bondat d'ajust, hem vist que la Deviança s'utilitza per a mesurar la discrepància en els MLG, podem expressar el seu càlcul com a suma de les desviacions unitàries $d_i = d(y_i, \hat{\mu}_i)$, és a dir:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d_i$$

El *residu de Deviança* es calcula:

$$r_D = sign(y_i - \mu_i)\sqrt{d_i}$$

es compleix que la suma del seu quadrat és igual a la Deviança, és a dir:

$$\sum_{i}^{n} r_D^2 = D$$

La nostra variable resposta serà el nombre total de sinistres (Y) que segueix una distribució de Poisson, és una distribució discreta utilitzada, en general, per a dades de comptatge.
Al capítol següent, particularitzarem l'anàlisi de models al cas d'una base de dades que tingui una variable resposta obtinguda al comptar successos. Concretament plantejarem dues problemàtiques que ens podem trobar quan treballem amb distribucions discretes: la sobredispersió i l'excés de zeros i com poder solucionar-les. Més endavant, aplicarem algunes d'aquestes alternatives a la nostra variable resposta. Per això, hem particularitzat els càlculs d'aquest capítol a la distribució de Poisson, es poden consultar a l'Annex A.2. De la mateixa manera, que es poden consultar la particularització a la distribució Binomial, necessària per aquells models que s'utilitzen per compensar l'excés de zeros i que veurem al proper capítol.

Capítol 3

Models per a dades de comptatge

Els MLG són una eina molt útil per modelitzar variables amb una distribució no normal i són més flexibles que els ML ja que permeten considerar una funció de l'esperança com a combinació lineal de les variables explicatives i no l'esperança en sí. Aquesta teoria es pot particularitzar al tipus de base de dades que tinguis en el teu problema, per exemple, quan disposem d'una variable resposta que pren valors enters no negatius que provenen de comptar i no d'una classificació, direm que tenim *dades de comptatge* (Count data) i, de fet, aquest és el nostre cas, ja que la variable resposta sorgeix de comptar el nombre de sinistres total (Y) que té cada assegurat. La distribució clàssica utilitzada per modelitzar aquests casos és la Poisson, distribució discreta uni-paramètrica que té una forta limitació que és la igualtat entre la variància i l'esperança.
En aquest capítol, veurem distribucions alternatives que, possiblement, aproximen millor les observacions i que compensen les restriccions a les que estem sotmesos quan utilitzem la distribució de Poisson (punt de partida en aquest tipus d'anàlisi). Més endavant, aplicarem aquests mètodes a la nostra base de dades ja que presentarà alguns dels problemes que estudiem en aquest capítol. Per realitzar aquest capítol vam utilitzar els llibres [20] i [21] entre altres materials esmentats a la bibliografia, com per exemple els articles [23] i [22].

3.1 Distribució de Poisson: sobredispersió i excés de zeros

La distribució de Poisson serà el punt de partida per modelitzar variables resposta provinents de comptar successos. Si recordem, ja vam particularitzar els càlculs dels MLG a aquesta distribució, en aquestes seccions treballarem concretament dos problemes que ens podem trobar quan treballem amb una Poisson: la sobredispersió i l'excés de zeros. La propietat més restrictiva de la distribució de Poisson és l'*equidispersió* que consisteix en la igualtat de la variància i l'esperança, això es deu a que la distribució està parametritzada en funció d'un únic paràmetre μ i, per tant, tots els seus moments han de dependre d'aquest coeficient. En particular, tenim:

$$\mathbb{E}(Y) = \mu$$
$$Var(Y) = \mu$$

Empíricament, el més habitual és trobar-se amb dades que tenen una variància empírica major que l'esperança mostral, aquest fenomen es coneix amb el nom de *sobredispersió* (overdispersion). Tenim dues possibilitats per determinar si la base de dades amb la que treballem presenta aquest fenomen:

1. Calcular un estimador del paràmetre de dispersió ϕ a partir de l'estadístic de Pearson Generalitzat χ^2_p. Aquest estadístic segueix una distribució chi-quadrat, és a dir, que la seva esperança serà igual als graus de llibertat i, a partir d'aquí podem obtenir $\hat{\phi}$:

$$\hat{\phi} = \frac{\chi^2_p}{n-p} \tag{3.1}$$

El valor del paràmetre de dispersió en la distribució de Poisson és 1. Si $\hat{\phi}$ té un valor més gran significa que tenim sobredispersió i si pel contrari, és més petit, tindrem *infradispersió* (underdisperion), és a dir, una variància inferior a l'esperança de la distribució, però aquest cas és poc habitual.

2. Una altra manera de determinar-la seria a partir d'un test d'hipòtesi. Sabem que en la distribució de Poisson la variància i l'esperança són iguals, per tant, si no es compleix aquesta condició podríem expressar l'esperança de la variable com:

$$Var(y|x) = \mu + \alpha g(\mu)$$

on, α és un paràmetre desconegut i $g(\cdot)$ una funció desconeguda.
De tal manera que si fem el test d'hipòtesi:

$$\begin{cases} H_0: & \alpha = 0 \\ H_1: & \alpha > 0 \end{cases}$$

i rebutgem la hipòtesi nul·la tindrem sobredispersió. Per determinar la infradispersió hauríem de canviar la hipòtesi alternativa per $H_1 : \alpha < 0$. Tal i com es pot veure a l'Annex A.3, tenim diverses possibilitats per adaptar la sobredispersió entre les quals hi ha considerar models de quasi-versemblança, o considerar el model Binomial Negativa que s'exposarà a continuació.
La *zero inflació* es dóna quan la variable resposta assoleix el valor 0 un alt nombre de vegades, és a dir, la probabilitat empírica del zero supera amb escreix la que estableix la distribució, en aquest cas, de Poisson. La manera més senzilla de determinar aquesta situació és fent una taula de freqüències de la variable resposta i comprovar que la freqüència del zero és superior a l'assignada per una Poisson amb valor esperat igual a la mitjana aritmètica.
Hi han dues alternatives molt semblants per compensar aquest problema: els Hurdle models i els zero-inflated models, només difereixen en la interpretació de la part nul·la de la variable resposta. A la subsecció 3.3 veurem en que consisteix cadascún d'ells.

3.2 Distribució Binomial Negativa

En aquesta secció ens centrarem en definir la Binomial Negativa per ser l'alternativa més habitual quan sorgeix el problema de la sobredispersió. La *Binomial Negativa* (Negative Binomial, BN) és una distribució discreta que, normalment, s'utilitza per a mesurar el nombre d'èxits d'un conjunt independent i idènticament distribuït d'assajos de Bernoulli, abans d'obtindre $\theta \in \mathbb{Z}^+$ fracassos. Aquesta definició requereix que el paràmetre θ sigui un enter positiu, ara bé la Binomia Negativa està definida també per a qualsevol $\theta \in \mathbb{R}$ malgrat que en aquest cas no admet la interpreació anterior. Pel que respecte a la

sobredispersió, ens interessa la Binomial Negativa en el seu espai de paràmetres més general. Hi han moltes parametritzacions de la seva funció de probabilitat, en aquest cas utilitzarem la següent:

$$f(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)y!}\frac{\mu^y\theta^\theta}{(\mu+\theta)^{y+\theta}} \tag{3.2}$$

on $\theta \in \mathbb{R} \setminus 0$ és un paràmetre de forma, $\mu > 0$ i $\Gamma(\cdot)$ és la funció Gamma. Amb aquesta parametrització, BN pertany a la família exponencial sempre que el paràmetre θ sigui conegut. En cas de que no sigui així, el paràmetre es podria estimar a partir de les observacions.
A la taula següent, resumirem com hem de seleccionar les funcions vistes en la família exponencial per a que l'equació (3.2) pertanyi a aquest conjunt assumint que coneixem el paràmetre θ:

Paràmetres	**Binomial Negativa**
Paràmetre de dispersió	$\phi = 1$
Funció acumulada:	$b(\theta) = 0$
$\boldsymbol{c(y;\theta)}$ **:**	$c(y;\theta) = \log \frac{\Gamma(y+\theta)}{\Gamma(\theta)y!(e^\mu+\theta)^{y+\theta}}$

Taula 3.1: Selecció de les funcions per a que la distribució Binomial Negativa pertanyi a la família exponencial

En aquest cas, l'esperança i la variància són iguals a:

$$\mathbb{E}(Y) = \mu$$
$$Var(Y) = \mu + \tfrac{\mu^2}{\theta}$$

i com podem veure, l'avantatge d'utilitzar aquesta distribució és que l'esperança coincideix amb la de la Poisson i, en canvi, la variància és major degut a que estem afegint un terme positiu.
Aquesta opció ens ajuda a adaptar la sobredispersió, però en alguns casos aquest fenomen ve donat per la gran quantitat de zeros que pren la variable resposta. Si es dóna aquest cas necessitem aplicar mètodes alternatius que abordarem a continuació.

3.3 Models per l'excés de zeros

3.3.1 Hurdle Models

El *Hurdle Model* és un model de regressió desenvolupat per augmentar la probabilitat de l'excés de zeros de la variable resposta. Aquest model consta de dues components: els zeros determinats per la probabilitat $f_1(\cdot)$, és a dir:

$$Pr(Y = 0) = f_1(0)$$

i els successos estrictament positius amb probabilitat:

$$f_2(y|y > 0) = \frac{f_2(y)}{1 - f_2(0)}$$

Es suposa que amb probabilitat $f_1(0)$ s'observarà el valor zero i amb probabilitat $1 - f_1(0)$ s'observarà un valor estrictament positiu d'una distribució $f_2(y)$. Així doncs el model Hurdle s'interpreta com una barreja d'una distribució desconeguda en zero i una distribució estrictament positiva. Tot plegat ens quedar

que la funció de probabilitat dels Hurdle models té la forma següent:

$$f_{hurdle}(y) = \begin{cases} f_1(0) & \text{si } y = 0, \\ (1 - f_1(0))\frac{f_2(y)}{1-f_2(0)} & \text{si } y \geq 1 \end{cases}$$

Per exemple, la funció de densitat del Hurdle model d'una distribució de Poisson tindrà la següent expressió:

$$Pr(Y = y) = \begin{cases} p & y = 0 \\ (1-p)\frac{\mu^y}{(e^\mu - 1)y!} & y \geq 1 \end{cases}$$

Hem de destacar que els els models utilitzats per estimar els zeros i els utilitzats per estimar els comptes positius no calen que siguin els mateixos. Això vol dir que es poden utilitzar diferents conjunts de variables explicatives en cada cas. En general, s'utilitzen models de regressió Binomial per a la variable binària que determina si el succés és zero o positiu, concretament utilitza la regressió logit, que apareix a la secció de la *Distribució Binomial* de l'Apèndix A.2, per estimar el valor p. Com a distribucions per a f_2 s'utilitzen la Poisson o la Binomial Negativa. Utilitzar la BN té l'avantatge que permet gestionar l'excés de zeros a la vegada que la sobredispersió, mentre que la Poisson només permet compensar l'excés de zeros.

Podríem pensar que en aquest cas serà més difícil aplicar els procediments dels MLG estudiats en el Capítol 2, però l'estimador de màxim versemblança té l'avantatge que permet maximitzar cada funció de forma independent.

L'equació de regressió de l'esperança utilitza el logaritme com a funció link canònic i té la següent forma:

$$\log \mu = x^T \beta + \log(1 - f_1(0)) - \log(1 - f_2(0))$$

L'equació per a la probabilitat és sovint la següent:

$$logit(p) = (\boldsymbol{X^*})^{\boldsymbol{T}}\boldsymbol{\beta}$$

on, com ja s'ha dit les matrius $\boldsymbol{X}$ i $\boldsymbol{X^*}$ poden ser diferents.

3.3.2 Zero-inflated Models

Els *zero-inflated* és un altre model de regressió utilitzat per gestionar les variables respostes que mostren més zeros dels esperats per la distribució clàssica que caldria suposar. Aquest mètode es diferència de l'anterior en què considera que hi ha dos tipus de zeros: els zeros reals i els zeros estructurals. Els *zeros estructurals* són aquells en que la probabilitat de valer zero és 1 i *els zeros reals* són aquells que, sota condicions normals, la variable resposta pren valor 0. L'inconvenient és que som incapaços de diferenciar el tipus de zero quan treballem la base de dades. Així que es consideren dues funcions de densitat: una per al procés binari $f_1(\cdot)$ utilitzada per determinar el tipus de zero i una altra $f_2(\cdot)$ que s'usa en el procés de comptatge. Tot plegat podríem expressar la funció de densitat del model zero-inflat de la següent manera:

$$f_{\text{zero infl}} = \begin{cases} f_1(0) + (1 - f_1(0))f_2(0) & y = 0 \\ (1 - f_1(0))f_2(y) & y \geq 1 \end{cases}$$

En general, es prenen les mateixes distribucions que hem vist en el Hurdle model: la regressió logit per a modelitzar la variable binària i la distribució de Poisson o Binomial Negativa per a la variable de comptatge. De la mateixa manera que abans, prendre la Binomial Negativa permet lidiar amb el problema

de l'excés de zeros i la sobredispersió simultàniament.
En aquest cas l'equació de regressió de l'esperança també utilitza el logaritme com a funció canònica de link, i té la següent expressió:

$$\mu = p \cdot 0 + (1-p)e^{x^T\beta}$$

La funció de densitat de probabilitat pel model Poisson zero inflat (ZIP) amb paràmetre μ és la següent:

$$Pr(Y=y) = \begin{cases} p + (1-p)e^{-\mu} & y = 0 \\ (1-p)\frac{e^{\mu}\mu^{y}}{y} & y \geq 1 \end{cases} \tag{3.3}$$

on μ és el paràmetre de Poisson.
La funció de densitat del model amb la distribució Binomial Negatiu zero inflat (ZINB) té la forma següent:

$$Pr(Y=y) = \begin{cases} p + (1-p)(1+\theta\mu)^{-1/\theta} & y = 0 \\ (1-p)\frac{\Gamma(y+\theta)}{\Gamma(\theta)y!}\frac{\mu^{y}\theta^{\theta}}{(\mu+\theta)^{y+\theta}} & y \geq 1 \end{cases}$$

on μ i θ són els paràmetres de la Binomial Negativa.
El pròxim capítol marcarà un punt d'inflexió en el nostre treball degut a que serà l'inici de la part aplicada. Començarem fent una anàlisi exploratòria de la base de dades i podrem comprovar que haurem de lidiar amb un dels dos problemes enunciats en aquest capítol concretament amb l'excés de zeros.

Capítol 4

Base de dades

Abans de modelitzar la variable resposta, hem de fer una anàlisi exploratòria de la base de dades i és al que ens dedicarem en aquest capítol. Això vol dir que s'estudiaria quines variables tenim, es mirarà si tenen valors anòmals o no, s'escolliran les variables predictores i es categoritzaran les que es considerin apropiades.

4.1 Descripció de la base de dades

La nostra base de dades ha sigut proporcionada per la companyia d'assegurances Catalana Occident com un fitxer $.sav$. Per aconseguir una mostra aleatòria simple s'han seleccionat aquelles pòlisses acabades en 5 i amb data final de la pòlissa fins a dia $30/09/2016$, en total obtenim 100.900 casos. Per cada fila, tenim 36 variables: sis de les quals defineixen la variable resposta i són: el nombre i cost de sinistres corporals, de sinistres culpa de l'assegurat i de sinistres culpa contrari. La trentena restant constitueixen el conjunt de variables explicatives. Si recordem, hem esmentat en més d'una ocasió que la finalitat d'un model és predir i explicar la variable resposta utilitzant el menor número de variables predictores possibles, per això i degut al gran volum de variables que tenim, el primer pas consistirà en seleccionar quines d'aquestes es coneix que tenen més influència en el número de sinistres.
En aquesta base de dades tenim el número i costos per a tres tipus de sinistres: *sinistres corporals*, aquells en que hi han hagut persones amb lesions físiques, *sinistres culpa de l'assegurat* (Y_1), com el seu nom indica són aquells en que el causant del sinistre ha sigut el conductor del vehicle assegurat i *sinistres de culpa contrària* (Y_2), que són aquells en que el causant és una persona diferent de l'assegurat. La suma dels dos darrers tipus constitueixen el *nombre total de sinistres RC* (Y) d'una companyia asseguradora.
Inicialment, la nostra intenció era prendre com a variable resposta Y_1, però degut a problemes que explicarem al Capítol 5, finalment hem decidit treballar amb la variable Y:

$$\begin{aligned} \boldsymbol{Y}_1: &\quad \text{nº sinistres culpa de l'assegurat} \\ \boldsymbol{Y}_2: &\quad \text{nº sinistres culpa contraria} \\ \boldsymbol{Y} = &\quad \boldsymbol{Y}_1 + \boldsymbol{Y}_2 \end{aligned}$$

Per a cada pòlissa tenim més d'una entrada a la base de dades, això es deu a que tenim en compte el seu historial des de que va entrar a la companyia i s'estudien tots els seus moviments (renovacions, anulacions, canvis en la tarifa...) tenint en compte el temps d'exposició de la pòlissa per any natural. A continuació

s'explica amb un exemple: considerem una persona que contracta una assegurança per al seu turisme el dia $01/04/2014$, el temps d'exposició d'aquesta pòlissa en aquell any serà $75,34\%$ de l'any. L'any 2015 es dividirà en dues situacions: el temps d'exposició des de l'1 de gener fins el dia en que la pòlissa es renovi (es renova el mateix dia que es contracta) i des del dia de la renovació fins al final de l'any, és a dir, el temps d'exposició serà un $24,66\%$ de la primera situació més $75,34\%$ de la segona i totes dues sumen el 100%. A la pràctica no és tan senzill perquè es tenen en compte tots els moviments i canvis que es fan a la pòlissa, com per exemple podria ser contractar noves garanties com els danys propis, la qual cosa voldria dir que aquesta pòlissa tindria més de dues situacions aquell any.

D'aquesta manera, ens trobem en la situació en que tenim molts casos però aquests no són independents perquè una mateixa pòlissa constitueix més d'una observació i aquestes observacions estaran condicionades per la manera de conduir de l'assegurat. Hem de destacar, que a la pràctica les companyies d'assegurances modelitzen el nombre de sinistres suposant que les observacions són independents, ja que d'aquesta manera tindran més volum de dades per aproximar la variable resposta i, a més, la manera de conduir d'una persona evoluciona amb la maduresa i l'experiència. Això sí, a l'hora d'analitzar els resultats, donaran més pes a les dades més recents perquè són les que ens indiquen com ha estat el comportament de l'assegurat últimament. Així que seguirem la mateixa metodologia que es fa servir en aquest camp. Hem de fixar un criteri que permeti definir quina és la població que es vol modelitzar i seleccionar les variables que es creu poden ser més significatives.

De les 30 variables que recull la companyia, n'hi ha algunes que clarament tenen poca importància en el número de sinistres d'un vehicle, com és el cas per exemple de les variables: *codi postal del prenedor*, *número de portes del vehicle* o *el valor dels accessoris* que s'afegeixen. De les 30 variables inicials considerem que les 17 que poden ser més influents són les que apareixen a la Taula 4.1:

Variables que caracteritzen el conductor		**Variables que caracteritzen el vehicle**	
$\boldsymbol{X_1}$	Edat del conductor del vehicle	$\boldsymbol{X_{10}}$	Tipus de vehicle
$\boldsymbol{X_2}$	Antiguitat del carnet del conductor (anys)	$\boldsymbol{X_{11}}$	Combustible del vehicle
$\boldsymbol{X_3}$	Coincidència del prenedor, conductor i propietari	$\boldsymbol{X_{12}}$	Antiguitat del vehicle (anys)
$\boldsymbol{X_4}$	Província de circulació	$\boldsymbol{X_{13}}$	Valor total del vehicle
$\boldsymbol{X_5}$	Any de la situació de la pòlissa	$\boldsymbol{X_{14}}$	Potència del vehicle
$\boldsymbol{X_6}$	Temps d'exposició de la pòlissa ($\%$ d'any)	$\boldsymbol{X_{15}}$	Pes del vehicle
$\boldsymbol{X_7}$	Historial sinistral del prenedor	$\boldsymbol{X_{16}}$	Cilindrada del vehicle
$\boldsymbol{X_8}$	Nombre de pòlisses d'automòbils a la companyia	$\boldsymbol{X_{17}}$	Quocient pes/potència
$\boldsymbol{X_9}$	Fraccionament del pagament		

Taula 4.1: Variables explicatives que considerem que més poden influir en el número total de sinistres. Classificades segons si serveixen per caracteritzar al conductor o al vehicle.

Per a fer la descriptiva, totes aquestes variables explicatives es convertiran en categòriques en el cas que no ho siguin, d'aquesta manera els gràfics seran molt més entenedors. La conversió de variables explicatives en categòriques en l'entorn de la sinistralitat en companyies d'assegurances és l'habitual. Les variables X_3, X_4, X_9, X_{10} i X_{11} ja estan categoritzades. La resta de variables es categoritzaran amb l'assessorament d'experts de la companyia.

A continuació, s'explica en detall quin és el tractament que s'ha fet a algunes d'aquestes variables.

Antiguitat del carnet del conductor (X_2)

La variable X_2 descriu l'experiència del conductor indicant els anys que fa que té el carnet. Destaca una categoria d'aquesta variable, la 999 que senyala quan el conductor del vehicle ha sigut assignat per una empresa ja que no és possible conèixer l'antiguitat del carnet perquè habitualment aquests automòbils son conduits per més d'una persona. El comportament d'un conductor quan condueix un vehicle propi o d'una societat pot ser diferent, no només perquè si el vehicle no és propi es pot tenir un comportament més arriscat, sinó també perquè els vehicles que corresponen a una empresa poden estar en risc de patir un sinistre un nombre més elevat d'hores, i això augmentaria la variabilitat en la variable resposta. Aquesta variabilitat no és clar que es pugui modelar a través d'una explicativa dicotòmica que indiqui si el vehicle és particular o no i, per això, s'ha decidit centrar aquest treball únicament en els vehicles particulars. A la Taula B.1 podem veure els resultats d'aquesta variable distribuïts segons el nombre total de sinistres.

Any de situació de la pòlissa (X_5)

Començarem analitzant la variable X_5, any de la situació. Aquesta variable és categòrica i té tres categories: 2014, 2015 i 2016. Per tal de disminuir la variabilitat de la base de dades, en aquest projecte ens centrarem en l'anàlisi de les pòlisses d'un únic any. L'objectiu és determinar quin és l'any amb menys pòlisses duplicades per aconseguir una mostra el més independent possible. En primer lloc, definim una variable indicadora que senyali si una pòlissa està duplicada o no i després fem una taula de contingència amb aquesta variable i la X_5. Podem consultar la Taula B.2 a l'Apèndix B.

A la Taula B.2 podem veure que el 2016 és l'any amb més casos primaris i un total de 34.268 pòlisses, escollirem aquest com a període de temps, ja que d'una banda és el que ens proporciona les dades més recents i el que, d'altra, té més casos primaris.

Tipus de vehicle (X_{10})

El tipus de vehicle és una variable classificada en 23 categories i amb un comportament molt diferent entre elles. El conjunt de *Turismes i derivats* constituït per turismes, monovolums, tot terrenys i furgonetes derivades de turismes és el col·lectiu amb més pòlisses i, per tant, el que més freqüència sinistral té. D'altra banda, les furgonetes no derivades de turisme o camions són col·lectius amb molta probabilitat de patir un sinistre degut a que acostumen a ser vehicles utilitzats per transportar matèries, és a dir, que passen gran quantitat de temps a la carretera. El grup de les motocicles té poca freqüència sinistral, però els sinistres acostumen a involucrar lesions físiques.

Farem una taula de contingències per veure com es distribueixen les pòlisses en els sinistres segons el col·lectiu al que pertanyen. La Taula B.3 mostra que el conjunt on es centren les pòlisses són els turismes i derivats, per aquest motiu aquesta població serà el nostre objecte d'estudi.

Finalment, aconseguim una selecció de 30.000 pòlisses i la variable X_{10} passa a estar classificada en 4 nivells: *turismes*, *tot terrenys*, *monovolums* i *furgonetes sí derivades de turisme*.

Després de fer totes aquestes seleccions ens quedem amb un total de 29.936 pòlisses, per cada pòlissa 16 variables explicatives, degut a que hem seleccionat com a any d'estudi el 2016, és a dir, tenim $29.936 * 16 = 478.976$ valors en la nostra base de dades.

4.2 Anàlisi exploratòria

En aquest apartat, d'una banda es realitzarà l'estadística descriptiva de les variables explicatives per tal de detectar si hi ha observacions que puguin considerar-se valors atípics, determinar si la variabilitat és acceptable, si hi ha correlació significativa entre parelles de variables, etc. D'altra banda, s'analitzarà la relació amb la variable resposta amb l'objectiu de determinar quines d'aquestes 17 variables seleccionades s'inclouran en el model multivariant inicial.
Per a les explicatives categòriques es realitzarà un diagrama de Pareto on, a l'eix de les $Y's$ principal, s'indicarà el nombre de pòlisses que tenim per cada categoria. El diagrama de Pareto s'acompanyarà per una poligonal a l'eix de les $Y's$ segundari que mostrarà el percentatge de freqüència sinistral del nombre total de sinistres.
La *freqüència sinistral* es defineix com el quocient del nombre total de sinistres de cada categoria dividit per la suma del temps d'exposició de cada pòlissa en l'any 2016, és a dir:

$$\%frec = \frac{\sum \text{nº sinistres}}{\sum \text{temps d'exposició}} * 100$$

Hem de recordar que cada pòlissa tindrà com a màxim un temps d'exposició del 75% perquè la data final més gran és del $30/09/2016$. Dividirem l'estudi segons si les variables caracteritzen al conductor o al vehicle.

4.2.1 Variables que caracteritzen al conductor habitual

Estudiarem primer aquelles variables que caracteritzen al conductor habitual del vehicle. Recordem que eren les variables $X_1, \ldots, X_9$, de les quals hem de descartar la variable X_5 per haver seleccionat l'any 2016 com a període de temps.

Edat del conductor del vehicle i Antiguitat del carnet del conductor ($\boldsymbol{X_1}$, $\boldsymbol{X_2}$)
Atès que un cop t'has tret el carnet B sovint ja no deixes de conduir, té sentit que les variables X_1 i X_2 estiguin altament correlacionades. La Figura B.1, situada a l'Annex B, conté la gràfica de dispersió d'aquestes dues variables amb la recta de regressió associada. L'edat mínima per treure's el carnet a Espanya són 18 anys, és a dir, que l'antiguitat del carnet no podrà ser mai superior a la diferència entre l'edat del conductor i 18, per aquest motiu es veu a la gràfica una segona recta determinada que té per equació $X_2 = X_1 - 18$. Aquesta recta diferència dos grups d'observacions: en blau els que estan per sobre i en vermell els que estan per sota. Les 147 observacions ressaltades en vermell s'han de descartar de la mostra per no tenir lògica amb la realitat, és a dir, indiquen conductors que es van treure el carnet amb una edat inferior a 18, no verifiquen la relació anterior. A la Figura B.2 podem veure com queda el gràfic definitiu, un cop descartades aquestes observacions.
Per quantificar el nivell de correlació entre les dues variables calcularem el *coeficient de correlació*, recordem que s'utilitza per mesurar la relació lineal entre dues variables quantitatives i es calcula:

$$r_{XY} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

Aquest coeficient pren valors a l'interval $[-1, 1]$, com més proper està el valor a 1 o -1, més lineal és la relació entre les variables i si $r = 0$ vol dir que no hi ha cap relació lineal entre elles.

La correlació entre aquestes dues variables és de $0,87$, clarament molt propera a 1, la qual cosa vol dir que només serà necessari incloure una de les dues en el model multivariant. Si s'incloguessin les dues tindríem un problema de multicolinearitat en el model. S'ha decidit considerar la variable X_2 i categoritzar-la en 6 classes: ≤ 4, $5-10$, $11-15$, $16-20$, $21-25$ i >25.
Tanmateix, hem de destacar, que en la modelització del preu de les tarifes d'una pòlissa s'han d'incloure totes dues variables per motius comercials. Si només incloguéssim, per exemple l'edat del conductor, beneficiaríem a aquelles persones que s'han tret el carnet més tard i que, com a conseqüència, tenen una alta probabilitat de patir accidents. Ara bé, per a la modelització que es portarà a terme en el següent capítol només es tindrà en compte X_2.
La variable X_2 és una variable quantitativa que indica els anys que fa que el conductor habitual té el carnet. Suposem que aquesta variable tindrà una contribució negativa en la probabilitat del nombre de sinistres, és a dir, la lògica ens dicta que com més experiència tinguis conduint, menor serà la probabilitat de que pateixis un sinistre. A la Figura B.3 situada a l'Apèndix B, podem veure que el major nombre d'assegurances està concentrat en la població que té més de 25 anys d'experiència (representa un 54,76% de la mostra) i que el comportament de la freqüència sinistral és tal i com havíem anticipat, la poligonal decreix progressivament fins que en l'última categoria s'estabilitza. Clarament la distribució de les assegurances en les diferents categories d'antiguitat del carnet no és simètrica.

Coincidència del prenedor, conductor i propietari (X_2)
En el negoci de les assegurances es distingeixen tres tipus de figures en una pòlissa: el *prenedor* és aquella persona física o entitat a nom de la qual va la pòlissa i que paga la seva prima, el *propietari* que és aquella persona física o societat a nom de qui esta el vehicle assegurat, i el *conductor habitual* que és la persona que condueix el vehicle assegurat assíduament. La variable X_3 determina si en una pòlissa aquestes tres figures coincideixen o no, és a dir, és una variable dicotòmica que val 1 quan totes tres figures són la mateixa i zero altrament. Està provat estadísticament que quan no coincideixen les tres figures, la probabilitat de tenir sinistres és superior, encara que de fet el més habitual és que coincideixin.
A la Figura B.4 de l'Apèndix B, podem veure que es verifica la nostra intuïció, és a dir, quan coincideixen les tres figures la freqüència és inferior al cas en que són diferents. El $67,6\%$ de les pòlisses estan concentrades en la categoria "Iguals", ja ens interessa que sigui així per ser la que millors resultats presenta.

Província de circulació (X_4)
Agruparem les províncies de la variable X_4 per comunitat autònoma, així tindrem una variable categòrica amb 17 categories en comptes de 50. A la Figura B.5 de l'Annex B hi apareix el diagrama de Pareto que determina el nombre de pòlisses per comunitat autònoma acompanyat d'una poligonal que senyala la seva freqüència sinistral. Es pot veure que la comunitat autònoma amb més nombre de pòlisses en termes absoluts és Catalunya i la de menys és La Rioja. Pel que respecte a la sinistralitat per a comunitats autònomes veiem que la comunitat autònoma amb menor sinistralitat és La Rioja i la de major és Melilla. Sorprèn que sigui Melilla la que té un $30,44\%$ de sinistralitat. Ara bé, en aquesta en aquesta comunitat autònoma només hi ha 203 pòlisses que és una mostra molt petita. Important observar que en la majoria de comunitats autònomes la sinistralitat es troba entre el 10% i 15%. Seria lògic pensar que les grans ciutats, seran els nuclis de més sinistralitat. Pel que respecta a les comunitats autònomes de Catalunya i Madrid, que són les que inclouen les dues ciutats més gran d'Espanya s'observa que a Madrid hi ha una

sinistralitat 50% més alta que a Catalunya.

Temps d'exposició de la pòlissa (X_6)

La variable X_6 descriu el temps d'exposició de cada pòlissa en l'any 2016. La lògica ens dicta pensar que com menys temps portis a la companyia més probabilitat de tenir sinistre tindràs, ja que quan portes molt temps en una asseguradora, al client l'interessa mantenir un bon historial per a que no li augmentin la prima i poder gaudir dels avantatges. Aquesta variable, com es podrà veure en el capítol següent, jugarà el paper de variable offset quan s'ajusti un model de Poisson. És a dir, serà una variable el coeficient de la qual no hem d'estimar perquè té sentit que sigui constant igual a 1. Clarament, la probabilitat de tenir sinistres d'un assegurat dependrà del temps que tingui contractada la pòlissa amb la companyia, no és el mateix tenir un client amb un sinistre en una pòlissa que fa 9 mesos que està contractada, que passat només un mes.

Aquesta variable l'hem classificada en 4 categories i a la Figura B.6 de l'Annex B podem veure, d'una banda, que el 40% de les pòlisses es concentren en el darrer nivell i, per l'altra, que la poligonal mostra una tendència decreixent, disminuint la freqüència conforme augmenta el temps d'exposició. Com hem descartat les observacions d'anys anteriors, no podem garantir si les pòlisses de l'última categoria són de l'últim any o si tenen més historial amb la companyia, el darrer cas explicaria que milloressin el comportament.

Historial de sinistralitat del prenedor (X_7)

Aquesta variable assigna un número al prenedor de la pòlissa segons el nombre i tipus de sinistres que ha tingut en els darrers 5 anys, essent 40 el de millor comportament i 250 el de pitjor. Té sentit que aquesta variable tingui una contribució positiva en la probabilitat de sinistres del conductor, és a dir, té sentit que l'assegurat tingui una freqüència major de sinistres com més alt sigui el número que t'assigna aquesta variable.

A la Figura B.7 podem veure que el $68,6\%$ de les dades estan localitzada a la primera categoria, precisament la de millor comportament. La poligonal mostra una tendència creixent del percentatge de freqüència alhora que augmenta X_7 tal com cavia esperar.

Nombre de pòlisses d'automòbils a la companyia (X_8)

Aquesta variable determinar la quantitat de pòlisses d'automòbils que té contractada cada client amb la companyia. A la Figura B.8 podem veure que està categoritzada, però al model del capítol següent figurarà com una variable quantitativa. Aquesta variable suposem que tindrà una contribució negativa respecte a la probabilitat de sofrir un sinistre. És a dir, té lògica pensar que una persona com més polisses d'automòbils tingui contractades amb la companyia, menys sinistres patirà ja que l'interessarà conservar les bonificacions i que no li augmentin les primes. A la Figura B.8 de l'apèndix es veu que la majoria de persones només tenen contractada una pòlissa i, tal i com hem anticipat, amb un pitjor comportament que la resta. El comportament de la poligonal coincideix amb l'anticipat, descriu una tendència decreixent conforme augmenta el nombre de pòlisses contractades pel client.

Fraccionament del pagament (X_9)

Aquesta variable descriu la modalitat de pagament de la prima de la pòlissa. Està categoritzada en 3

categories: anual, semestral i trimestral. Té sentit pensar que la freqüència de les categories augmentara, conforme augmentin el nombre de pagaments. Hi ha diverses raons per fraccionar una prima: que aquesta sigui molt cara, la comoditat de poder-la pagar en diversos cops o no tenir solvència financera per fer front als pagaments. Un altre motiu, podria ser aprofitar el fraccionament per fer el primer pagament per a poder gaudir de gairebé tot l'any assegurat, declarant molts sinistres i al final canviar de companyia abans de finalitzar el contracte.
A la Figura B.9 es pot veure que, afortunadament, el $77,2\%$ de pòlisses fan pagaments anuals. La poligonal mostra el comportament que anticipàvem, creix de forma monòtona conforme augmenta el nombre de pagaments.

4.2.2 Variables que caracteritzen el vehicle assegurat

Ja hem fet l'anàlisi exploratòria de totes aquelles dades que conformaran el perfil del conductor habitual del vehicle assegurat, ara estudiarem aquelles que caracteritzen el tipus de vehicle que s'assegura. Recordem que són les variables $X_{10}, \ldots, X_{17}$.

Tipus de vehicle ($\boldsymbol{X_{10}}$)
A la secció de la descripció de la base de dades, ja hem transformat la variable X_{10} en una categòrica de només quatre categories: turismes, monovolums, tot terrenys i furgonetes derivades de turisme.
A la Figura B.10 de l'Apèndix B podem veure que el col·lectiu de turismes concentra el major nombre de pòlisses, de fet representa el 70% d'assegurances. Els tot terreny són els vehicles amb menor freqüència de sinistres, mentre que les furgonetes derivades són les que la tenen més alta, la qual cosa té sentit perquè aquest tipus de vehicle passa moltes hores a la carretera.

Combustible del vehicle ($\boldsymbol{X_{11}}$)
La variable X_{11} indica el combustible del vehicle assegurat. Està distribuïda en 3 categories: dièsel, benzina i resta. En general, els vehicles dièsel tenen més sinistralitat que els de benzina, ja que són vehicles pensats per a persones que utilitzen molt el cotxe.
A l'Annex B podem trobar la Figura B.11, on es pot veure que la majoria de pòlisses tenen com a combustible dièsel o benzina i, de fet, el $66,14\%$ estan localitzades en la primera categoria. La poligonal mostra més sinistralitat en vehicles de dièsel que de benzina. La freqüència de la darrera categoria és molt elevada, però la dada no és gaire fiable perquè només es tenen 102 assegurats en aquesta categoria.

Antiguitat del vehicle($\boldsymbol{X_{12}}$)
Aquesta variable ens proporciona el nombre d'anys del vehicle des del dia la seva matriculació. Aquesta variable té un rang de $0 - 45$ i esta agrupada en 7 categories segons la seva freqüència sinistral. El comportament d'aquesta variable podria tenir diverses vessants: d'una banda com més anys té un cotxe, més problemes mecànics li poden sorgir i, com a conseqüència més probabilitat de tenir un sinistre, però d'altra banda, el conductor coneix millor el vehicle i anteposarà mantenir el seu historial net a l'asseguradora abans que declarar sinistres lleus.
A la Figura B.12 podem veure que el més habitual és tenir assegurats amb vehicles d'entre 7 i 15 anys i bastant estrany aquells que superen els 20. Podem veure que la poligonal mostra una influència negativa

de la variable i més relacionada amb la segona explicació. A més, un altra explicació podria ser que aquesta variable es confongués amb X_2, podríem pensar que una persona que s'acaba de treure el carnet es comprarà un cotxe nou que anirà sent més antic conforme augmenti la seva antiguitat del carnet, és a dir, que totes dues variables podrien arribar a tenir el mateix comportament. A la Figura B.13 podem veure que el núvol de punts no descriu una relació lineal entre les variables, és a dir, no sembla que hi hagi cap correlació, la qual cosa queda confirmada pel coeficient de correlació que és $0,097$, molt proper a zero. El fet que no estiguin fortament correlacionades ens portarà a incloure ambdues variables en la modelització que es portarà a terme en el capítol següent.

Valor total del vehicle (X_{13})

La variable X_{13} ens aporta el preu final de vehicle sumant els accessoris que se li afegeixen. Aquesta variable està categoritzada en 12 categories segons si tenen la mateixa freqüència de sinistres. Aquesta variable pot tenir dos tipus de comportament: d'una banda, pot tenir més sinsitres perquè com més val un vehicle més potència té, però d'altra també inclou avenços tecnològics que el poden fer més segur.
A la Figura B.14 podem observar que el $45,5\%$ dels vehicles assegurats tenen un valor d'entre 15.000 i 25.000 euros, mentre que només un $0,6\%$ estan a la categoria de més de 90.000, no interessa tenir cotxes de molt valor degut a que com major sigui aquest, més cares seran les reparacions dels desperfectes. Esperàvem que la freqüència sinistral augmentés a mesura que augmenta el valor del vehicle i, de fet, això és el que hem obtingut, una poligonal amb una tendència creixent, però de forma irregular. Hem de destacar, que els resultats de les categories amb menys grandària mostral són poc fiables.

Valor total del vehicle, Potència del vehicle i Cilindrada del vehicle (X_{13}, X_{14} i X_{16})

Quan parlem de *potencia d'un vehicle* ens referim a la velocitat màxima que aquest pot assolir, bàsicament és la relació entre el treball realitzat i el temps en que es realitza, es mesura en cavalls. D'altra banda, la *cilindrada d'un vehicle* és la suma del volum útil de tots els cilindres d'un motor i es mesura amb centímetres cúbics. Ambdues definicions estan relacionades: si augmenta la cilindrada d'un vehicle, augmentarà el volum dels seus pistons i la seva empenta relativa, aquesta empenta es transformarà en treball útil i, com a conseqüència augmentarà la potència. Degut a això, té sentit suposar que les variables X_{14} i X_{16} puguin estar altament correlacionades. A més, també podria tenir sentit pensar que un vehicle com més car sigui el seu cost, més potència tindrà, això voldria dir que hi hauria correlació entre les variables X_{13} i X_{14}.
A la Figura B.15 de l'apèndix podem observar que la potencia i la cilindrada tenen una relació lineal positiva, de fet, el seu coeficient de correlació és $0,758$, bastant proper a 1, així que només inclourem una en el model multivariant del proper capítol. De fet, en les companyies asseguradores s'utilitza normalment la cilindrada per modelitzar les motocicletes i la potencia per la resta de col·lectiu. En canvi, la variable valor total no mostra cap correlació amb les altres dues, el coeficient de correlació és $0,67$. En els gràfics que apareix el valor total, esta ressaltada en vermell una observació que té un valor del vehicle molt alt (és un *outlier*). Per tal de disminuir la variabilitat de les dades i atès que aquesta observació no és habitual, descartarem aquesta pòlissa. A la Figura B.16 hem repetit la matriu d'scatter plot anterior, però sense l'outlier. Podem observar que la potència i cilindrada mantenen el seu comportament lineal amb un coeficient de correlació $0,760$. En canvi, el valor total ara mostra una relació lineal amb la potència i cilindrada. El coeficient de correlació de valor total amb potència és $0,885$ i de valor total amb cilindrada

és $0,792$. Per tant, inclourem només el valor total en l'anàlisi del proper capítol i no es posarà la cilindrada ni la potència per evitar el problema de multicolinearitat.

Quocient pes/potència (X_{17})
A la literatura científica s'ha vist que a a vegades enlloc d'incloure les variables pes i potència per separat en el model multivariant, s'inclou una única variable corresponent al quocient d'ambdues. En aquest apartat definirem i descriurem la variable X_{17} definida com el quocient de les variables X_{15} i X_{14}:

$$\boldsymbol{X_{17}} = \frac{X_{15}}{\boldsymbol{X_{14}}}$$

Els vehicles amb un valor de X_{17} petit tenen més probabilitats de patir accidents més greus, un valor petit ens diu que el pes del vehicle és inferior a la potencia que ha de suportar. Aquesta classe de vehicles es diu que estan *sobrepotenciats* i tendeixen a bolcar quan tenen accidents. Esperem que aquesta variable tingui una influència negativa en el fet de patir o no un sinistre, és a dir, com major sigui el valor del coeficient menor hauria de ser la seva freqüència.
A la Figura B.17 podem veure que la majoria de pòlisses tenen vehicles amb un quocient entre 10 i 11, la qual cosa vol dir que el pes del vehicle està compres entre 10 i 11 vegades la seva potència. El comportament de la freqüència sinistral no s'assembla al que suposàvem d'entrada, només coincideix fins al valor 14, després la freqüència torna a créixer, això podria ser per la poca grandària mostral d'algunes de les categories. Hem de destacar que abans de fer el càlcul del quocient, hem descartat aquelles pòlisses amb pes 0, degut a que són situacions en que no tenim informació del pes del vehicle. Això ha suposat descartar 57 entrades.

4.3 Base de dades final

En aquest capítol hem fet molts canvis en la nostra base, així que en aquesta secció resumirem quins hem fet i, finalment, a quina base de dades hem arribat, que és la que s'utilitzarà en la modelització del capítol següent.
La nostra base de dades original constava de 100.900 pòlisses. Els criteris inicials que es van establir per per a determinar la població objecte d'estudi vàren ser els següents:

- **Antiguitat del carnet del conductor:** a la base de dades original hi havia 64 pòlisses amb una societat com a prenedor. Aquestes observacions han sigut descartades ja que el comportament d'un conductor amb un vehicle propi és molt diferent del que condueix un cotxe d'empresa.

- **Any de situació de la pòlissa:** a la base de dades inicial hi ha un gran volum d'observacions, així que per reduir variabilitat, decidim seleccionar únicament les pòlisses de l'any 2016. D'aquesta manera, acabem tenint 34.268 casos i una variable explicativa menys.

- **Tipus de vehicle:** continua sent un volum de dades considerable, i com no tots els vehicles es comporten de la mateixa manera, decidim treballar amb el col·lectiu que representa la majoria de les pòlisses: Turismes i Derivats. Així, aquesta variable queda categoritzada en 4 classes: turismes,

monovolums, tot terrenys i furgonetes derivades. Per tant, al final d'aquestes decisions tenim un total de 29.936 pòlisses.

A continuació mostrarem aquells criteris que hem hagut d'establir un cop feta l'anàlisi exploratòria de cada variable explicativa degut a que les dades no tenien un sentit real:

- **Antiguitat del carnet del conductor:** al fer l'scatter plot de l'antiguitat del carnet i l'edat del conductor, vam poder veure que hi havia un grup d'observacions que no verificaven la relació que existeix entre aquestes dues variables. Per llei, una persona es pot treure el carnet a partir dels 18 anys, això vol dir que l'antiguitat del carnet no pot ser superior a la diferència entre l'edat del conductor i 18. Les 147 dades que no verificaven aquesta desigualtat van ser descartades.

- **Pes del vehicle:** en algunes de les pòlisses figurava un 0 com a pes del vehicle, això és degut a que no tenim la informació exacta d'aquesta característica. Vam descartar aquestes observacions, atès que no feien possible el càlcul de la variable X_{17}.

Un cop aplicats tots aquests criteris, obtenim que la mostra final està constituïda per 29.732 pòlisses.
Per cada fila tenim 13 variables, contant la variable resposta, és a dir, en total tenim 386.516 valors a la base de dades.
A continuació, farem un breu llistat de les variables independents que es consideraran per a la modelització inicial del capítol següent. Entre parèntesi hi figura el nom que tenen en la nostra base de dades:

- $\boldsymbol{X_2}$ **(edacar):** es mesura en anys i té un rang de $0-68$, correspon a l'antiguitat del carnet de conduir. Hem vist, que la influència d'aquesta variable és negativa, és a dir, com més anys d'experiència té un conductor, menys accidents tindrà degut a que es guanya destresa en al conducció.

- $\boldsymbol{X_3}$ **(mismafig):** aquesta variable indica si les tres figures d'una pòlissa (prenedor, propietari i conductor) són iguals o diferents. Hem vist que les pòlisses en que no coincideixen les tres figures tenen una freqüència de sinistres major, ja que el client pretén combinar diferents persones amb bons comportaments per a pagar menys prima. El més normal és que coincideixin les tres figures.

- $\boldsymbol{X_4}$ **(procir:)** correspon a la província de circulació del vehicle, però l'hem categoritzada per comunitat autònoma. Catalunya és la comunitat autònoma amb major número de pòlisses contractades. Els resultats de la freqüència són irregulars però cal esmentar que en algunes comunitats autònomes hi ha un nombre molt petit de pòlisses.

- $\boldsymbol{X_6}$ **(tempexp):** aquesta variable indica el percentatge de temps d'exposició de la pòlissa en l'any 2016. Ja hem comentat, que es prendrà com una variable offset ja que hem de comparar el nombre de sinistres en proporció al temps que porti l'assegurat en la companyia aquell any.

- $\boldsymbol{X_7}$ **(hpsrct):** aquesta variable té una influència negativa en la sinistralitat, perquè denota l'historial de la sinistralitat del prenedor de millor a pitjor comportament. Conforme pitjor historial té, més freqüència.

- $\boldsymbol{X_8}$ **(autpol):** aquesta variable indica el nombre de pòlisses d'automòbils contractades per l'assegurat. Hem vist que la poligonal decreix, però de forma no gaire accentuada. Esperàvem veure una poligonal decreixent, ja que com més pòlisses tens, més motius per a conservar el bon historial i no perdre bonificacions. Podria ser que en la modelització ens sortís no significativa.

- $\boldsymbol{X_9}$ **(pagpol):** aquesta variable esta categoritzada en tres classes: els que fa un únic pagament anual, els que el paguen semestralment o trimestralment. Hem vist que la freqüència dels sinistres augmenta conforme augmenten el nombre de pagaments, això és degut a que les persones que el fraccionen podrien fer-ho amb la intenció d'estar assegurat gaire bé tot l'any i quan arriba el segon pagament marxen a una altra companyia.

- $\boldsymbol{X_{11}}$ **(comveh):** en l'anàlisi exploratòria d'aquesta variable hem vist que els vehicles que utilitzen dièsel tenen una freqüència sinistral més alta que els que utilitzen benzina. No podem dir res, dels vehicles que utilitzen un altre combustible per la falta de grandària mostral. La diferència entre la sinistralitat dels dièsel i benzina vindrà o no confirmada amb el model multivariant en el cas que la variable surti estadísticament significativa.

- $\boldsymbol{X_{12}}$ **(antveh):** el rang d'aquesta variable és $0 - 44$ anys, suposem que com més anys té un vehicle més freqüència sinistral tindrà degut a que la seva mecànica està més antiquada i més deteriorada. D'altra banda, aquesta variable té certa relació amb l'experiència del conductor, degut a que vehicles amb més anys estaran conduits per persones amb més experiència i, per tant, aquests conductors s'anticiparan a possibles contratemps.

- $\boldsymbol{X_{13}}$ **(valtot):** aquesta variable indica el preu del vehicle, el seu valor màxim és de $137.554,49$ euros. Suposem que com més car sigui el vehicle, més possibilitat de declarar un sinistre tindrà, degut a que aquesta variable està altament correlacionada amb la potència. Ara bé, no forçosament ha de ser així, perquè també pot ser que els conductors de cotxes cars siguin molt més prudents.

- $\boldsymbol{X_{17}}$ **(quocient):** correspon al quocient pes/potència, té un rang de $2 - 64$. L'evolució de la seva freqüència sinistral no era tal i com esperàvem, comença decreixent fins a la quarta categoria però després tornar a créixer. Algunes de les seves categories, com per exemple la 14, $18 - 22$ i ≥ 23, tenen poca grandària mostral la qual cosa podria fer els nostres resultats poc fiables.

La variable resposta serà analitzada i modelitzada al següent capítol, i es veurà quines d'aquestes variables surten que tenen una influència significativa en la resposta.

Capítol 5

Modelització de la variable nombre de sinistres

Com hem esmentat a la introducció, aplicar la teoria estudiada a dades reals és complicat perquè sorgeixen molts problemes, la realitat és sovint força diferent del que dicta la teoria i ho hem pogut comprovar de primera mà. En un principi, anàvem a modelitzar el nombre de sinistres culpa de l' assegurat, variable categoritzada en 3 classes: 0, 1 i 2. Per fer-ho, vam realitzar una petita selecció de la base de dades aplicant els criteris exposats al Capítol 4 i vam centrant-nos en la comunitat autònoma de Catalunya i en el turisme com a tipus de vehicle. El problema al modelitzar-la va ser que no obteníem variables significatives. El motiu pel qual res ens sortia significatiu era perquè el 98% de les observacions tenien zero sinistres culpa assegurat. Clarament estàvem davant d'un excés de zeros, però tan gran que el programa utilitzava el model nul per aproximar la variable resposta, ja que la considerava com si fos idènticament zero.
Per tal d'anar més enllà, vam decidir canviar la variable resposta al nombre total de sinistres (suma dels que són culpa de l'assegurat i dels que són culpa contraria). D'aquesta manera augmentem el nombre de pòlisses amb sinistres. Al modelitzar aquesta variable amb la mateixa selecció que abans, vam aconseguir obtenir alguna variable significativa, però continuaven sent massa poques.
Finalment, vam decidir ampliar la selecció de la població per modelar, i no considerar només Catalunya i turismes sinó totes les comunitats autònomes i el col·lectiu de Turismes i Derivats. Així vam acabar realitzant la selecció resumida a la secció *Base Final* del capítol anterior: hem seleccionat l'any 2016, hem pres els Turismes i Derivats com a tipus de vehicle i només hem considerat aquelles vehicles en que el prenedor és una persona física. D'aquesta manera, obtenim una base de dades amb 29.732 entrades.
En les següents seccions, mostrarem els procediments i els resultats obtinguts de la modelització de la variable resposta. Per realitzar aquest capítol, ens hem basat en el format de dos articles que modelitzaven les seves variables respostes aplicant els models del Capítol 3. El primer de tots és [22], aplicat al camp de la medicina i molt útil ja que exposa les llibreries i funcions necessàries per realitzar la modelització d'aquest capítol. El segon article important és [23] que té l'avantatge que aplica alguns dels models del Capítol 3 al camp dels sinistres d'automòbils amb la diferència que ho fa per la garantia Danys Propis.
En la propera secció farem una 5.2 ajustem un ML a la variable descriptiva de la variable resposta que volem modelitzar. En la secció 5.2 ajustarem un ML a la variable Y prèviament transformada. Això és el que es feia abans de l'existència dels MLG i per això hem considerat pertinent fe-ho. La secció 5.3 s'ajust aun model ZIP. A la darrera secció es porta a terme un estudi comparatiu dels MLG ajustats.

5.1 Descriptiva de la variable resposta

La variable resposta és el nombre total de sinistres i esta classificada en 3 categories: 0, 1 i > 2 sinistres. Inicialment assumirem que aquesta variable segueix una distribució de Poisson. Almenys 1.687 pòlisses (el $5,68\%$ de la mostra) han tingut algun sinistre i les 28.045 restants cap (un $94,33\%$ de la mostra). Al Capítol 3 vam veure que quan treballem amb variables que segueixen una distribució discreta se'ns poden presentar dos problemes: l'excés de zeros i la sobredispersió. A la Taula 5.1 podem veure la distribució dels sinistres en les tres categories:

	0 sinistres	**1 sinistre**	**$>$ 2 sinistres**
Nombre de pòlisses	28.045	1.590	97

Taula 5.1: Distribució de les pòlisses segons el nombre de sinistres totals que s'han declarat

En aquesta taula podem observar clarament que hi ha un excés de zeros a la nostra base de dades.
Una altra manera de comprovar-ho, i la més habitual, és realitzar un histograma de la variable resposta i així podem veure gràficament com es distribueixen les observacions. La Figura 5.1 mostra les freqüències del nombre total de sinistres.

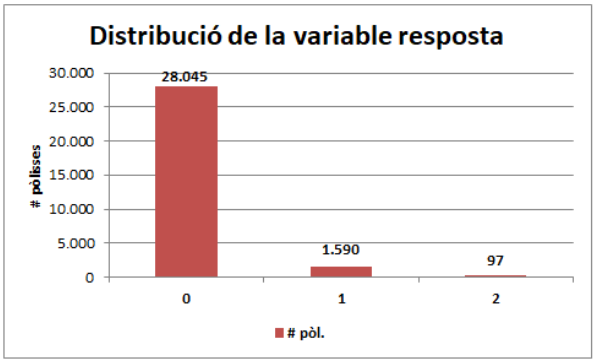

Figura 5.1: diagrama de Pareto de la variable resposta nombre total de sinistres

Es pot veure, com la categoria de zero sinistres concentra el major nombre de pòlisses. Per tant, això ens demostra que dels dos problemes que hem explicat al Capítol 3 almenys l'excés de zeros sí que el tenim. I, per tant, caldrà assumir un model ZIP o un ZINB. Per a veure si les dades tenen o no sobredispersió caldrà ajustar un model de Poisson i estimar el paràmetre de dispersió (ϕ).

5.2 Model Lineal amb resposta transformada

Abans de disposar de la teoria dels MLG, quan es tenia una variable resposta que no seguia una distribució normal s'utilitzaven procediments alternatius que aprofitaven la teoria dels ML. S'intentava transformar la variable resposta de forma que fos aproximadament normal i després s'aplicava l'anàlisi dels ML explicat al Capítol 1. Quan la variable seguia una distribució de Poisson, les transformacions suggerides eren dues:

$$Y_1 = \sqrt{Y + \frac{3}{8}}$$
$$Y_2 = Y^{2/3}$$

A l'Annex C.1 s'ha modelat seguint la teoria dels ML, la variables resposta Y amb el llistat de variables explicatives explicat a la Secció 4.3 del capítol anterior. D'aquests models lineals, el que més ens interessa analitzar és el coeficient R^2, que si recordem era el que mesurava com s'ajusta el model a les observacions. Amb la primera transformació obtenim un $R^2 = 0,02034$ i amb la segona transformació un $R^2 = 0,02036$, valors molt semblants i molt petits, la qual cosa vol dir que molta informació de la variable resposta queda sense explicar amb aquesta selecció de variables. En ambdós models també obtenim uns residus que no poden assumir-se normals i d'aquí que passem directament als MLG.

5.3 Regressió de Poisson

En aquesta secció assumirem que la variable nombre total de sinistres segueix una distribució de Poisson amb un nombre esperat que depèn de les variables explicatives. Com a funció link prendrem el seu link canònic, que si recordem era la funció logaritme.
El model inicial que hem considerat és el següent:

$$\begin{aligned}\log \mu \quad &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \log X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \\ &+ \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{17} X_{17}\end{aligned} \tag{5.1}$$

D'aquestes variables seleccionades, no totes surten significatives, no ho són les explicatives X_{10}, X_{11}, X_{13} i X_{17}. Per saber-ho, estem aplicant el test d'hipòtesi que vam veure a la secció *Hipòtesi sobre els paràmetres* del Capítol 1. Descartarem d'una en una aquestes variables segons el seu p-valor, començarem per aquella que el tingui més alt. Així que l'ordre en que les he suprimit és: X_{10}, X_{17} i X_{11}. Atès que quan eliminem alguna variable deixem informació de la variable resposta sense explicar, podem aconseguir que altres variables esdevinguin significatives, com en aquest cas la variable X_{13} que inicialment no ho era però ha passat a ser-ho quan suprimim les altres. Així que el model final amb el que treballarem consta de 9 variables explicatives, de les quals: X_6 és un offset, X_3, X_4 i X_9 són factors i les 5 restants són variables quantitatives. L'equació d'aquest model final és:

$$\log \mu = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \log X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{12} X_{12} + \beta_{13} X_{13} \tag{5.2}$$

A l'Annex C.2 podem veure el codi R utilitzat per ajustar aquest model així com els resultats obtinguts.

Interpretació dels resultats

La interpretació dels coeficients de les variables explicatives dependrà de si són factors o variables quantitatives. Per modelitzar un factor, l'R estableix la primera categoria com a base i després compara la resta de categories amb aquesta. De manera que la significació o no dels coeficients de cadascuna de les categories, indica si la categoria en qüestió és o no diferent de la que s'ha pres com a base.
Passem a continuació a interpretar els coeficients obtinguts en el model de Poisson:

- **Antiguitat del carnet del conductor (X_2)**
  La variable X_2 es pren com a variable quantitativa. Per interpretar aquesta variable, considerarem que tenim un client amb un determinat perfil definit per les explicatives. El nombre de sinistres

esperat per aquest individu és:

$$\log \mu = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{12} X_{12} + \beta_{13} X_{13} \quad (5.3)$$

Si mantenim tot el seu perfil igual, excepte la variable X_2 que l'augmentem un any. El nombre esperat de sinistres μ^* complirà:

$$\begin{aligned} \log \mu^* &= \beta_0 + \beta_2 (X_2 + 1) + \beta_3 X_3 + \beta_4 X_4 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{12} X_{12} + \beta_{13} X_{13} \\ &= \log \mu + \beta_2 \end{aligned} \quad (5.4)$$

Així doncs,

$$\log \mu^* = \log \mu + \beta_2 \Leftrightarrow \log \frac{\mu^*}{\mu} = \beta_2 \Leftrightarrow \mu^* = e^{\beta_2} \mu = e^{-0,071} \mu = 0,93 * \mu$$

i podem veure que, en aquest cas, el nombre de sinistres quan el client té un any més d'experiència i es mantenen la resta de condicions igual queda multiplicat per $0,93$. És a dir, disminuirà el nombre de sinistres esperat tal com cabia esperar.

- **Coincidència del prenedor, conductor i propietari (X_3)**
  Aquesta variable té dues categories: 0 quan les figures són diferents i 1 quan són iguals. El programa estableix la categoria 0 com a base i ha determinat que el comportament entre les dues categories és diferent i, per tant, la variable és significativa. Per analitzar els resultats d'aquesta variable, considerem dos clients (client 1 i 2) amb el mateix perfil i el mateix tipus de vehicle que només difereixen en la variable X_3, és a dir, en la pòlissa del client 1 no coincideixen el prenedor, el conductor i el propietari mentre que en la del client 2 sí.
  L'equació d'ambdós clients serà diferent, en el cas del client 1 el valor de X_3 s'estableix inicial a zero per ser la base, és a dir:

$$\log \mu = \beta_0 + \beta_2 X_2 + \beta_4 X_4 + \log X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{12} X_{12} + \beta_{13} X_{13} \quad (5.5)$$

  i en el cas del client 2 la variable X_3 pren el valor 1, amb la qual cosa:

$$\log \mu^* = \beta_0 + \beta_2 X_2 + \beta_3 + \beta_4 X_4 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{12} X_{12} + \beta_{13} X_{13} \quad (5.6)$$

  Atès que tenen el mateix perfil de client i de vehicle excepte per la variable X_3, la segona equació (5.6) es pot escriure en funció de la primera (5.5) i es té que:

$$\log \mu^* = \log \mu + \beta_3 \Leftrightarrow \log \frac{\mu^*}{\mu} = \beta_3 \Leftrightarrow \mu^* = e^{\beta_3} \mu$$

  i, si substituïm el paràmetre β_3 pel valor que ens estima el model tenim:

$$\mu^* = e^{-0,1808} \mu = 0,8346 * \mu$$

  Així doncs la freqüència de sinistres quan les figures coincideixen serà inferior a quan no ho fan, és a dir, les pòlisses localitzades a la categoria *Iguals* tindran millor comportament que les de *Diferents*. El nombre de sinistres del client 2 és el resultat de multiplicar el nombre esperat quan les figures no coincideixen per $0,835$.

- **Província de circulació (X_4)**
  La variable X_4 està categoritzada en 18 classes perquè recordem que l'agrupem per comunitat autònoma, i es pren com a base la comunitat autònoma de Catalunya. Les comunitats que tenen un comportament significativament diferent del de Catalunya són: Galícia, Madrid i Melilla, mentre que la resta tindran un comportament molt similar. Això es pot veure en la significació dels coeficients que apareix en la sortida del codi de l'Apèndix C.2. És a dir, si suposem que tenim 4 clients que tenen les mateixes característiques del conductor habitual i del vehicle, a excepció de la seva comunitat autònoma de circulació: el client 1 circularà per Catalunya, el client 2 per Galícia, el client 3 per Madrid i el client 4 per Melilla. De la mateixa manera que abans, podem expressar les $\mu's$ dels clients 2, 3 i 4 en funció de la del client 1 i obtenim:

$$\mu_{Galícia} = e^{\beta_4}\mu_{Cat} = e^{0,223}\mu_{Cat} = 1,25 * \mu_{Cat}$$
$$\mu_{Madrid} = e^{\beta_4}\mu_{Cat} = e^{0,309}\mu_{Cat} = 1,36 * \mu_{Cat}$$
$$\mu_{Melilla} = e^{\beta_4}\mu_{Cat} = e^{0,949}\mu_{Cat} = 2,58 * \mu_{Cat}$$

  El valor del paràmetre dels clients $2, 3$ i 4 és positiu, la qual cosa vol dir que les seves freqüències sinistrals seran superiors a les de Catalunya. És a dir, la probabilitat de tenir un sinistre dels clients 2, 3 i 4 serà superior a la del client 1. Hem de destacar que la comunitat autònoma de Melilla, que és la que té un nombre esperat de sinistres més gran, tenia una grandària mostral molt petita, concretament de 203 pòlisses, és a dir, podria ser que les diferències fossin degudes a tenir poques observacions.

- **Historial sinistral del prenedor (X_7)**
  La variable X_7 també és quantitativa, així que tindrà una interpretació similar a la de X_2. En aquest cas, considerarem dues situacions diferents d'un client: la primera situació serà la seva situació actual de la pòlissa i la segona situació serà mantenir totes la variables de la situació 1 iguals menys la variable X_7 que augmentarem en una unitat el que vol dir que el nou conductor té un pitjor comportament que el de la situació 1. Podem expressar la μ^* de la segona situació en funció de la μ de la situació 1 de la manera següent:

$$\mu^* = e^{\beta_7}\mu = e^{0,007}\mu = 1,007 * \mu$$

  Així doncs, podem veure que si mantenim totes les variables iguals menys la X_7 que augmenta una unitat, la freqüència sinistral del client augmentara molt poc perquè el nombre de sinistres queda multiplicat per 1 si arrodonim. Ara bé, això no va senyit amb que la variable sigui significativa ja que quan X_7 l'augmentem en força més d'una unitat llavors sí que les diferències poden ser importants.

- **Nombre de pòlisses (X_8)**
  La variable X_8 també és quantitativa. Seguirem el mateix procediment que en la resta de variables quantitatives. Considerarem la situació actual de la pòlissa i després una segona situació que augmentarà una unitat el nombre de pòlisses i mantindrà la resta de variables iguals. Tal i com hem fet abans, podem expressar la μ^* de la segona situació en funció de la μ de la primera:

$$\mu^* = e^{\beta_8}\mu = e^{-0,077}\mu = 0,463 * \mu$$

Podem veure que d'una situació a l'altra el nombre de sinistres queda multiplicat per $0,463$. La qual cosa vol dir que conforme augmenten el nombre de pòlisses a la companyia disminueix el nombre de sinistres. Això té sentit amb l'anàlisi exploratòria que havíem fet al Capítol 4 d'aquesta variable. També lliga amb el fet que si un té més pòlisses a la companyia intenta ser més prudent i no tenir sinistres per tal que se li mantiguin les condicions.

- **Fraccionament del pagament (X_9)**
  La variable X_9 està categoritzada en 3 categories: anual, semestral i trimestral i el programa pren com a categoria base l'anual. En aquest cas, per fer l'anàlisi considerem 3 clients amb el mateix perfil del conductor habitual i del vehicle que només difereixen en la variable X_9. El client 1 no fracciona la seva prima, el client 2 la pagarà de forma semestral i el client 3 trimestralment. Com la resta de variables són iguals podem expressar les $\mu's$ dels clients 2 i 3 en funció de la del client 1:

$$\begin{aligned} &\text{client 2} \quad \mu^* = e^{\beta_9}\mu = e^{0,245}\mu = 1,28 * \mu \\ &\text{client 3} \quad \mu^* = e^{\beta_9}\mu = e^{0,355}\mu = 1,43 * \mu \end{aligned}$$

  Podem veure que els paràmetres β_9 dels clients 2 i 3 són positius, la qual cosa vol dir que els comportaments de les pòlisses que paguen la prima semestral o trimestralment són pitjors que les que ho fan anualment. Això es tradueix amb que els clients que fraccionen la seva prima trimestralment tenen un nombre esperat de sinistres que és $1,43$ vegades més gran que els que no la fraccionen.

- **Antiguitat del vehicle (X_{12})**
  Atès que la variable X_{12} és quantitativa, considerarem la situació actual de la pòlissa i una segona situació que mantingui el perfil del conductor habitual i del vehicle i augmentarem un any la seva antiguitat. Podem expressar la μ^* de la segona situació en funció de la μ de la primera:

$$\mu^* = e^{\beta_8}\mu = e^{-0,084}\mu = 0,91 * \mu$$

  En aquest cas, el nombre de sinistres d'una situació a una altre queda multiplicat per $0,91$. És a dir, que si augmenta l'antiguitat del vehicle aleshores es reduirà lleugerament el nombre de sinistres esperat.

- **Valor total del vehicle (X_{13})**
  Aquesta és al darrera variable inclosa en el model i també és quantitativa. Com sempre, considerarem dues situacions de la pòlissa, la pòlissa actual i una segona situació que mantingui tota la resta de variables del model iguals menys la variable X_{13} que l'incrementarem en una unitat. Expressem la μ^* de la segona situació en funció de la μ de la primera situació:

$$\mu^* = e^{\beta_{13}}\mu = e^{0,042}\mu = 1,04 * \mu$$

  El nombre de sinistres esperat quedarà multiplicat per $1,04$, és a dir, que augmentarà a mesura que augmenti el valor total del vehicle. Això potser degut a que els cotxes amb un valor alt són capaços d'anar a una velocitat més gran, el que comporta més risc de patir un sinistre.

5.4 Model de Poisson zero inflat

En aquesta secció ajustarem la mateixa variable resposta assumint que la seva distribució és una ZIP enlloc d'una Poisson. Això permetrà estimar un paràmetre més que clarament augmentarà la probabilitat de tenir zero sinistres.
Com vam explicar al capítol 3, les variables explicatives que s'utilitzen per estimar la probabilitat del zero i el nombre de sinistres no tenen per què ser iguals, així que prendrem per a l'estimació de la probabilitat el model constant i per al nombre de sinistres els que hem pres en la regressió de Poisson. El codi i els resultats per al model ZIP es poden trobar a l'apèndix C.3
Obtenim que la constant és significativa en la regressió logit, així que podem estimar la probabilitat p del model ZIP (3.3) aplicant la següent equació:

$$logit\left(\frac{e^p}{1-e^p}\right) = \hat{\beta}_0 \Leftrightarrow \hat{p} = \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} = \frac{e^{-0,959}}{1+e^{-0,959}} = 0,2771$$

Aquest valor ens està dient que el $27,71\%$ de les persones mai reporten un sinistre. Mentre que el 72% restant té un comportament Poisson respecte el nombre de sinistres, i alguns d'ells també reportaran zero sinistres. Com que hem utilitzat el mateix model pe a la part de Poisson que en la regressió de Poisson, no farem la interpretació dels paràmetres ja que no difereix gaire de la realitzada en la secció anterior.
Ara que ja disposem d'un model que ens adapta l'excés de zeros, som capaços de verificar si a la base de dades hi ha sobredispersió o no. Recordem, que es feia mitjançant l'estimació del paràmetre de dispersió ϕ:

$$\tilde{\phi} = \frac{\chi^2_{n-p}}{n-p} = \frac{34.758,3}{29.732-10} = 1.17$$

Com podem veure, l'estimació de ϕ és molt propera a 1, la qual cosa vol dir que no hi ha sobredispersió i, per tant, no serà necessari aproximar les nostres dades mitjançant la distribució Binomial Negativa zero inflada.

5.5 Comparació d'ambdós models:

A la Taula 5.2 mostrem les mesures habituals de bondat d'ajust que s'utilitzen per a comparar models

	# paràmetres	loglikelihood	AIC	Deviança	χ^2
Poisson	9	-6.511	13.074	9.589	$1,17$
ZIP	10	-6.506	13.066	–	$1,14$

Taula 5.2: Taula comparativa de la regressió de Poisson i ZIP

Tenint en compte aquests resultats podem dir que el model de Poisson zero inflat s'ajusta millor a les nostres observacions, perquè té un AIC i un χ^2 més petits. També té un valor de la log-versemblança lleugerament més gran. Amb el model ZIP tenim un paràmetre més que és la constant que utilitzem per augmentar la probabilitat del zero que al sortir significativament diferent de zero ens corrobora també que el model ZIP és més apropiat que el de Poisson.
A l'Annex C.4 podem consultar els gràfics dels residus versus valors predits d'ambdós models realitzats (Figures C.1 i C.2). No es habitual realitzar aquestes gràfiques quan s'apliquen MLG aquí no passa com als ML que han de provindre d'una distribució Normal. Ara bé, els hem realitzat perquè també ens aporten

informació. Quan es treballa amb una variable de Poisson que té moltes categories, s'espera veure uns residus més normalitzats. En el nostre cas, tenim una variable Poisson amb només 3 categories, de fet, són les tres corbes que es poden contemplar als gràfics. En totes dues gràfiques podem veure que els residus semblen descriure una funció exponencial, la qual cosa tindria sentit degut a que és la inversa de la funció logaritme, link canònic que s'esta utilitzant en ambdós models. A la Figura C.1 podem veure que aproximem millor els sinistres de la categoria 0 corresponent a la primera corba que els de les altres categories. Les altres dues categories tenen residus més alts. Té sentit que sigui així, ja que la grandària mostral de la categoria 0 és molt major que la de resta de categories. En el model ZIP, podem veure que els residus tenen una forma molt similar als residus de la regressió de Poisson, però per a la categoria zero (la que conté el major nombre de dades) els seus valors són molt més propers a zero. De fet, els residus de la categoria de 0 sinistres valen tots gairebé 0. Hi ha alguns residus però de les altres dues categories que surten superiors als obtinguts pel model de Poisson.

Conclusions

A l'inici del treball havíem dividit en dos tipus els nostres objectius, segons si eren teòric o pràctics. Podem dir, que hem assolit tots els nostres objectius teòrics que ens havíem plantejat. Així que aquestes conclusions les centrarem en la part pràctica del treball que engloba els Capítols 4 i 5.
Les conclusions principals de forma esquemàtica són les següents:

- Amb la nostra base de dades no es podia modelar la variable *nombre de sinistres culpa de l'assegurat* perquè aquesta era pràcticament igual a zero. Per això s'ha modelitzat la variable *nombre de sinistres total*.

- La primera anàlisi portada a terme ha estat una regressió de Poisson on han sortit significatives les variables: antiguitat del carnet (X_2), coincidència del prenedor, propietari i conductor (X_3), província de circulació (X_4), historial de sinistralitat del prenedor (X_7), nombre de pòlisses d'automòbils contractades amb la companyia (X_8), fraccionament del pagament (X_9) antiguitat del vehicle (X_{12}) i valor total del vehicle (X_{13}). I no significatives les variables: tipus de vehicle (X_{10}), tipus de combustible (X_{11}) i quocient pes/potencia (X_{17}). La variable temps d'exposició (X_6) s'ha considerat un variable offset. De les significatives X_4, X_9 i X_{13} augmenten el nombre de sinistres a l'augmentar i X_2, X_3, X_7 i X_{12} disminueixen el nombre de sinistres a l'augmentar.

- Degut a que la variable resposta té un 94% de zeros, s'ha considerat apropiat millorar el model assumint que la resposta segueix una ZIP enlloc d'una Poisson. Efectuant aquest ajust s'obté que la probabilitat p estimada del model ZIP és $\hat{p} = 0,277$ que és significativament diferent de zero. Les explicatives que surten significatives per la part corresponent al comptatge són les mateixes que pel model de Poisson, amb interpretacions semblants ja que els paràmetres no canvien gaire.

- Finalment s'han comparat els dos models mitjançant els estadístics habituals de Bondat d'Ajust: AIC, χ^2 i log-versemblança arribant a la conclusió que el model ZIP és significativament millor que el model de Poisson.

- No s'ha ajustat el model Binomial Negatiu ja que en el nostre cas no ha estat necessari perquè el paràmetre de dispersió del model ZIP sortia pràcticament igual a la unitat.

Apèndix A

Apèndix de conceptes de models

En aquest apèndix hem incorporat, tots els conceptes, exemples o teoremes que consideràvem útils de fer un cop d'ull, però no estrictament necessaris per al desenvolupament del projecte. Distingirem les seccions segons el capítol al que fan referència.

A.1 El Model Lineal

Recordem, que aquest és el capítol 1 del treball. Així que en aquesta secció podrem trobar totes aquelles referències de l'apèndix que hem realitzat en aquest mateix capítol.

A.1.1 El mètode general de la Màxima Versemblança

En aquesta secció, explicarem el mètode general de l'estimació de la màxima versemblança, que després al capítol particularitzarem al cas de la normal.
Considerem una mostra independent i idènticament distribuïda $y_1, \ldots, y_n$ de la variable aleatòria Y amb funció de densitat $f(y_i, \theta)$. La *funció de densitat conjunta* de la variable Y serà:

$$f(Y;\theta) = f(y_1;\theta)f(y_2;\theta)\cdots f(y_n;\theta)$$

I la *funció de versemblança*:

$$L(\theta) = f(Y;\theta)$$

Com hemt dit, l'estimador de màxima versemblança, $\hat{\theta}_{ML}$, serà aquell que maximitzi la funció $L(\theta)$. En general, no es treballa amb aquesta funció directament, si no que s'utilitza la funció log-versemblant. Com que el logaritme és una funció creixent, l'estimador que maximitzi aquesta funció serà el que maximitzi la nostra funció original. La *funció log-versemblant* serà:

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} log(f(y_i, \theta))$$

Per calcular el màxim, derivem respecte el paràmetre θ i igualem a 0:

$$\frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{n} \frac{1}{f(y_i;\theta)} \frac{\partial}{\partial\theta} f(y_i, \theta) = 0$$

Aïllant θ d'aquesta expressió obtindríem l'estimador. Per comprovar que aquest és el màxim hauríem de calcular la segona derivada i comprovar que aquesta avaluada en $\hat{\theta}_{ML}$ és negativa.

A.1.2 El teorema de Gauss-Markov

Abans d'enunciar el Teorema de Gauss-Markov hem de definir el concepte de funció paramètrica estimable. Una funció ψ és *paramètrica* si la podem expressar com una combinació lineal:

$$\psi = a_1\beta_1 + \ldots + a_k\beta_k$$

i és *estimable* si existeix un estadístic $\hat{\psi}$ que puguem expressar com una combinació lineal de les observacions $y_1, \ldots, y_n$, és a dir:

$$\hat{\psi} = b_1 y_1 + \ldots + b_n y_n$$

i tal que verifiqui $\mathbb{E}(\hat{\psi}) = \psi$. Aquesta condició s'imposa per què d'aquesta manera l'estadístic $\hat{\psi}$ serà un estimador no esbiaixat.
Ara ja disposem de tots els conceptes per enunciar el teorema:

Teorema A.1.1 *Si tenim una funció paramètrica estimable $\psi = \boldsymbol{a}^T\boldsymbol{\beta}$ i l'estadístic $\hat{\beta}$ és un estimador LS de β, aleshores:*

(a) L'estimador $\hat{\psi} = \boldsymbol{a}^T\boldsymbol{\beta}$ és únic.

(b) $\hat{\psi} = \boldsymbol{a}^T\boldsymbol{\beta}$ és l'estimador de variància mínima dins del conjunt dels estimador no esbiaixats de ψ.

(c) Si es compleixen les hipòtesi de normalitat aleshores l'estimador de $\boldsymbol{\beta}$ calculat pels mínims quadrats i per la màxima versemblança és el mateix.

A.1.3 Inferència sobre els estimadors

Un cop calculats els estimadors al capítol 1, podem veure que tots ells depenen linealment de les observacions. Ara pot ser útil, conèixer quines distribucions segueixen cadascun i això ho podem arribar a determinar a partir de les hipòtesi dels residus, exposades en la secció de l'estructura del ML.
En l'expressió de β (1.5) es pot veure que l'estimador depèn linealment de la variable resposta i sabent la distribució de $\boldsymbol{Y}$ (1.2) podem dir que $\boldsymbol{\beta}$ seguirà una distribució normal multivariant:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2) \tag{A.1}$$

a continuació veiem com es calculen els paràmetres de la distribució:

$$\begin{array}{lll}
\text{Esperança:} & \mathbb{E}(\hat{\boldsymbol{\beta}}) & = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta} \\
\text{Variància:} & Var(\hat{\boldsymbol{\beta}}) & = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Var(\boldsymbol{Y})((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)^T = \\
 & & = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}_n\sigma^2((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)^T = \\
 & & = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2
\end{array}$$

A partir de l'equació (1.10) es pot veure que l'estimador $\hat{\boldsymbol{Y}}$ segueix una normal multivariant:

$$\hat{\boldsymbol{Y}} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{P}\sigma^2) \tag{A.2}$$

amb esperança i variància:

$$\begin{aligned} &\text{Esperança:} \quad \boldsymbol{E}(\hat{Y}) = \boldsymbol{P}\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} \\ &\text{Variància:} \quad Var(\hat{\boldsymbol{Y}}) = \boldsymbol{X}Var(\hat{\boldsymbol{\beta}})\boldsymbol{X}^T = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2 = \boldsymbol{P}\sigma^2 \end{aligned}$$

Finalment, seguint la mateixa idea que en els dos casos anteriors tenim que l'estimador dels residus $\boldsymbol{e}$ també segueix una normal multivariant:

$$\boldsymbol{e} \sim N(\boldsymbol{0}, (\boldsymbol{I} - \boldsymbol{P})\sigma^2)$$

de paràmetres:

$$\begin{aligned} &\text{Esperança:} \quad \mathbb{E}(\boldsymbol{e}) = (\boldsymbol{I}_n - \boldsymbol{P})\mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{X} - \boldsymbol{P}\boldsymbol{X})\boldsymbol{\beta} = (\boldsymbol{X} - \boldsymbol{X})\boldsymbol{\beta} = 0 \\ &\text{Variància:} \quad Var(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{P})\sigma^2 \end{aligned}$$

A.1.4 Bondat d'ajust del model

Linealitat:
A la Figura A.1 podem veure un scatter plot de Y vs X que mostra la relació de linealitat entre les dues variables.

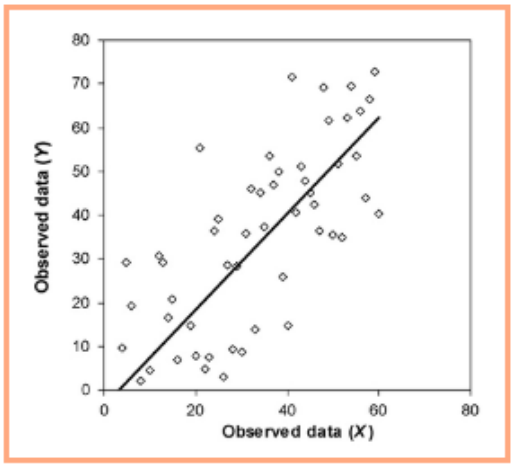

Figura A.1: Scatter-plot de Y vs X, on es mostra la relació lineal entre les variables

Homocedasticitat:
En la Figura A.2, situada a continuació, es mostren dos scatter-plots de $\hat{e}$ vs $\hat{y}$: en el de la dreta veiem com la variabilitat de les observacions augmenta quan creix la variable resposta (heterocedasticitat), mentre que en el gràfic de l'esquerra només es veu un núvol de punts que no segueix cap patró (homocedasticitat).

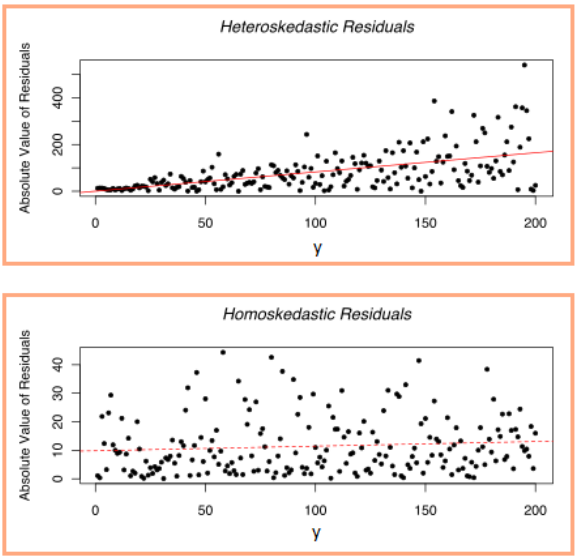

Figura A.2: Scatter-plot de e vs $\hat{y}$: a sota es mostra l'homocedasticitat (igualtat de variàncies) i a dalt, el contrari, heterocedasticitat.

Normalitat dels residus:
En la Figura A.3 s'ensenyen dos qq-plots: en el de l'esquerra es veu una clara relació lineal mentre que les cues del de la dreta s'allunyen de la recta.

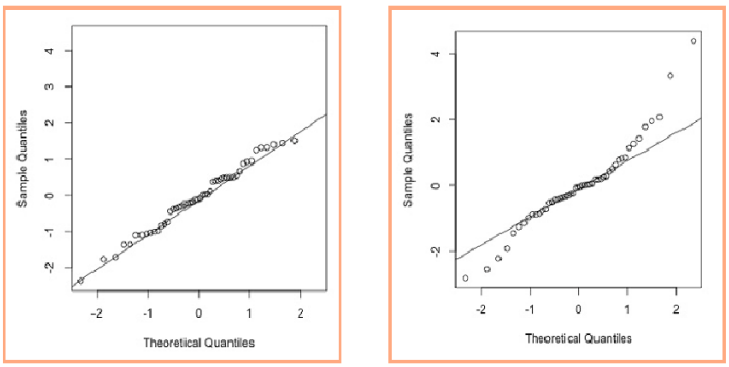

Figura A.3: Gràfics qq-plot de l'estimació dels errors: a l'esquerra s'observa la relació lineal demanada, mentre que en el de la dreta les cues s'allunyen de la recta desitjada.

A.1.5 Cas particular: la regressió lineal simple

Un cop queda plantejat com es procedeix en el tractament d'un model lineal, per finalitzar el capítol, repetirem els càlculs per al cas de la *regressió lineal simple* que és l'exemple més senzill de model lineal que està format per una única variable predictora.
L'equació dels ML en aquest cas és:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

la seva representació matricial és:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Tal i com hem vist, calcularem els estimadors dels paràmetres amb l'equació normal (1.4) que en aquest cas segueix la següent expressió:

$$\begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix} \Rightarrow$$

Aïllant el vector de paràmetres $\boldsymbol{\beta}$:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{n\sum(x_i - \bar{x})^2} \begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix}$$

obtenim que prenen com a valor:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i \bar{x})^2} = \frac{\sum y_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \end{aligned}$$

Ja hem vist, que el següent pas seria estimar la variable resposta, l'error i la suma de quadrats:

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ e_i &= y_i - \hat{y}_i \\ SSE &= \sum e_i^2 = \sum(y_i - \hat{y})^2 \end{aligned}$$

L'estimador centrat de la variància σ^2 és:

$$s^2 = \frac{\sum(y_i - \hat{y_i})^2}{n - 2}$$

El vector de paràmetres $\boldsymbol{\beta}$ segueix una distribució normal multivariant de paràmetres:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2} & \frac{-\sigma^2 \bar{x}}{\sum(x_i - \bar{x})^2} \\ \frac{-\sigma^2 \bar{x}}{\sum(x_i - \bar{x})^2} & \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \end{pmatrix} = \boldsymbol{\Sigma} \right)$$

Els ic dels paràmetres β_0 i β_1 són:

$$\begin{aligned} \beta_0 &\in \left(\hat{\beta}_0 - t_{n-2,\alpha} s \sqrt{\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}}, \hat{\beta}_0 + t_{n-2,\alpha} s \sqrt{\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}} \right) \\ &56 \\ \beta_1 &\in \left(\hat{\beta}_1 - t_{n-2,\alpha} \frac{s}{\sqrt{\sum(x_i - \bar{x}}}, \hat{\beta}_1 - t_{n-2,\alpha} \frac{s}{\sqrt{\sum(x_i - \bar{x}}} \right) \end{aligned}$$

En aquest exemple també es pot aplicar el test d'hipòtesi sobre el paràmetre $\boldsymbol{\beta}$, encara que no tingui gran interès perquè llavors estaríem en la situació del model nul. El test d'hipòtesi que es realitza consisteix en:

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$$

És a dir, si es compleix la hipòtesi nul·la tindrem que l'equació del model serà $y_i = \beta_0 + \epsilon_i$.

A.2 Els Models Lineals Generalitzats

En aquest apartat trobarem totes aquelles notacions esmentades en el capítol 2.

A.2.1 Cas particular: el Model Lineal

Hem definit els MLG com una extensió dels ML, explicats en el capítol anterior. En aquesta secció només volem comprovar que realment és així i ho farem veient que els ML són un cas particular i que, per tant, tenen definides les tres components que caracteritzen els MLG. A continuació, determinarem cadascuna d'aquestes components:

- **Component aleatòria**: ja sabem que la variable resposta d'un MLG ha de pertànyer a la família exponencial. La variable depenent dels ML segueix una distribució Normal, així que el primer pas consistiria en comprovar que aquesta distribució forma part d'aquesta família, és a dir, que la seva funció de densitat té la forma de l'equació (2.1).
  La funció de densitat de la distribució Normal és:
  $$f(y;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$
  Si fem el canvi $\bar{f}(y;\mu) = e^{\log f(y;\mu)}$ en la funció anterior tindrem:
  $$\begin{aligned} \bar{f}(y;\theta) &= e^{-\frac{y^2}{2\sigma^2}+\frac{y\theta}{\sigma^2}-\frac{\theta^2}{2\sigma^2}-\frac{1}{2}\log 2\pi\sigma^2} \\ &= e^{\frac{y\theta-\frac{\theta^2}{2}}{\sigma^2}-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+\log 2\pi\sigma^2\right)} \end{aligned}$$
  que té la mateixa forma que la funció de densitat de la família exponencial. A la taula següent, mostrem com hem de prendre les funcions per a que així sigui:

Paràmetres	**Normal**
Paràmetre de dispersió:	$\phi = \sigma^2$
Funció acumulada:	$b(\theta) = \frac{\theta^2}{2}$
$\boldsymbol{c(y;\theta)}$	$c(y;\theta) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log 2\pi\sigma^2\right)$
Esperança, $\mathbb{E}(Y;\theta)$:	$\mu(\theta) = \theta$

Taula A.1: Taula que mostra com prendre els paràmetres en la funció de densitat d'una Normal per a que tingui la mateixa forma que una família exponencial

- **Component sistemàtica:** hem vist que aquesta component engloba la matriu de condicions experimentals i el vector de paràmetres, en el cas del ML és exactament igual, és a dir:
  $$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np-1} \end{pmatrix}; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$
- **funció link:** com hem dit, aquesta funció relaciona l'esperança amb la component sistemàtica del model i en el cas dels ML aquesta funció és la identitat. Atès que els errors verifiquen $\boldsymbol{\epsilon} \sim N(0,\sigma^2)$, l'esperança de la variable resposta depèn linealment de les condicions experimentals:
  $$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{Y}) = \mathbb{E}(\boldsymbol{X\beta} + \boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{X\beta}) = \boldsymbol{X\beta} \Rightarrow \boldsymbol{\mu} = \boldsymbol{X\beta}$$

Per tant, els ML són un cas particular dels MLG, ja que hem sigut capaços de determinar cadascuna de les tres components que els caracteritzen.

A.2.2 Distribució de Poisson

En aquesta secció de l'apèndix veurem que la distribució de Poisson pertany a la família exponencial, és a dir, que podem aplicar la teoria estudiada al capítol 2 dels MLG.
Hem de determinar quines son les tres components dels MLG en aquesta distribució:

- **Component aleatòria:** considerem que la variable resposta segueix una *distribució Poisson* de paràmetre λ, recordem que és una distribució de probabilitat discreta utilitzada per modelitzar successos en que ens interessa determinar la probabilitat de que un cert número d'esdeveniments succeeixin en un cert període de temps.
  La seva funció de densitat té l'expressió:

$$f(y,\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

  Per veure que pertany a la família exponencial considerarem el canvi $\bar{f}(y,\theta) = e^{\log f(y,\theta)}$, la funció de densitat ens queda:

$$\bar{f}(y,\lambda) = e^{\log\left(\frac{\lambda^y e^{-\lambda}}{y!}\right)} = e^{y\log\lambda - \lambda - \log y!}$$

  Imposem la condició de que $\theta = \log\lambda$ per la forma que ha de tenir la funció de densitat per a pertànyer a la família exponencial (2.1), aleshores:

$$\bar{f}(y,\theta) = e^{y\theta - e^{\theta} - \log y!}$$

  A la taula següent, resumirem com hem de seleccionar les funcions per tal que la funció de densitat de la Poisson tingui la mateixa forma que la funció de la família exponencial:

Paràmetres	**Poisson**
Paràmetre de dispersió	$\phi = 1$
Funció acumulada:	$b(\theta) = e^{\theta}$
$\boldsymbol{c(y;\theta)}$ **:**	$c(y;\theta) = -\log y!$

Taula A.2: Selecció dels paràmetres per a que la distribució de Poisson pertanyi a la família exponencial

  En el cas d'aquesta distribució, l'esperança és igual al paràmetre λ, és a dir:

$$\mu = \mathbb{E}(Y) = \lambda = e^{\theta}$$

  Recordem que la distribució de Poisson té un característica particular i és que la seva variància i esperança són iguals, és a dir:

$$Var(\mu) = \mu$$

  En el pròxim capítol, veurem que a la pràctica una variable amb una distribució de Poisson no verifica aquesta condició, es necessiten tècniques alternatives per arreglar-ho.

- **Component sistemàtica:** com hem repetit en diverses ocasions, aquesta component engloba la matriu de condicions experimentals $\boldsymbol{X}$ i el vector de paràmetres $\boldsymbol{\beta}$.

- **Funció link:** la funció canònica que relaciona l'esperança amb la component sistemàtica és el logaritme. La raó és que la variable resposta té com a domini $\mathbb{Z}^+$, però la funció μ està definida a tot $\mathbb{R}$ i l'única funció que relaciona tots els reals amb l'eix positiu és el logaritme.

Per tant, podem aplicar la teoria dels MLG a una variable resposta amb distribució Poisson perquè hem sigut capaços de determinar cadascuna de les components que els caracteritzen.
Passarem directament a veure com es calculen els coeficients de bondat d'ajust i els residus perquè no es pot calcular quina forma hauria de tenir l'estimador de la variable resposta sense una base de dades.
Per tant, a continuació enumerarem cadascun dels estadístics de bondat d'ajust i dels residus i els calcularem per a aquesta distribució concreta:

- **Deviança:** recordem que era dues vegades la diferencia entre la funció log-versemblant del model complert i la del model de p paràmetres, en el cas d'una Poisson l'expressió que segueix és:

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}(y_i \log y_i/\hat{\mu_i} - (y_i - \hat{\mu}_i))$$

  Com ja hem vist, el paràmetre de dispersió d'aquesta distribució és 1 la qual cosa vol dir que la Deviança i la Deviança escalada són la mateixa.
  Per aquesta distribució aquest estadístic es coneix com G^2 *de Bishop*. A la pràctica, habitualment només es considera el primer terme per brevetat. Està totalment justificat perquè sempre que el model fitat tingui un terme constant, la suma d'aquest segon terme serà idènticament zero.

- **Estadístic generalitzat de Pearson:** en el cas de la distribució de Poisson coincideix amb l'estadístic X^2 de Pearson i té la següent expressió:

$$X^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}}$$

  Ja hem especificat abans, aquest estadístic tendeix a una distribució χ^2 asimptòticament.

- **Residu de Pearson:** recordem que és bàsicament la definició de residu escalada per l'estimació de la Deviança estàndard de Y. Rep aquest nom, perquè en el cas de la distribució de Poisson és el signe de l'arrel quadrada de l'estadístic anterior. O sigui, que la seva expressió queda:

$$r_P = \frac{y - \mu}{\sqrt{\mu}}$$

- **Residu de la Deviança:** com hem explicat a la teoria, cada unitat contribueix una quantitat d_i a la Deviança i la formula del residu s'expressa com un múltiple de l'arrel d'aquesta quantitat. En el cas de Poisson té la següent forma:

$$r_D = sign(y - \mu)\sqrt{2}\left(y\log(y/\mu) - y + \mu\right)^{1/2}$$

A.2.3 Distribució Binomial

En aquesta secció seguirem el mateix procediment que en l'anterior, començarem determinant les tres components dels MLG en cas de que la variable resposta tingui distribució Binomial i després particularitzarem els càlculs a aquesta situació.

- **Component aleatòria:** suposem que la variable resposta segueix una distribució Binomial, distribució discreta que mesura el nombre d'èxits en una seqüència de n assajos independents de Bernoulli amb probabilitat p d'èxit, la seva funció de probabilitat és:

$$f(y,p) = \begin{pmatrix} n \\ y \end{pmatrix} p^y (1-p)^{n-y}$$

Si fem el mateix canvi que anteriorment serà més fàcil determinar les funcions per a comprovar que pertany a la família exponencial, si recordem era $\tilde{f}(y,p) = e^{\log f(y,p)}$:

$$\tilde{f}(y,p) = e^{\log \begin{pmatrix} n \\ y \end{pmatrix} p^y (1-p)^{n-y}} = e^{\log \begin{pmatrix} n \\ y \end{pmatrix} + y \log \frac{p}{1-p} + n \log 1-p}$$

En aquest cas, imposem la condició $\pi = \log \frac{p}{1-p}$ per a que la funció segueixi la mateixa forma que l'equació (2.1):

$$\tilde{f}(y,\pi) = e^{\frac{y\pi - n \log 1 + e^{\pi}}{1/n} + \log \begin{pmatrix} n \\ y \end{pmatrix}}$$

A la taula següent mostrem un resum de la selecció de les funcions:

Paràmetres	**Binomial**
Paràmetre de dispersió	$\phi = \frac{1}{n}$
Funció acumulada:	$b(\pi) = \log 1 + e^{\pi}$
$\boldsymbol{c(y;\pi)}$ **:**	$c(y;\pi) = \log \begin{pmatrix} n \\ y \end{pmatrix}$

Taula A.3: Selecció dels paràmetres per a que la distribució Binomial pertanyi a la família exponencial

L'esperança de la distribució Binomial en funció dels nous paràmetres ens quedaria:

$$\mathbb{E}(Y) = \mu = \frac{e^{pi}}{1 + e^{\pi}}$$

i la funció variància depenent del paràmetre μ té l'expressió:

$$Var(Y) = \mu(1-\mu)$$

- **Component sistemàtica:** aquesta component és exactament la mateixa que en la Poisson, $\boldsymbol{\eta} = \boldsymbol{X\beta}$.

- **Funció link:** la funció link de la distribució Binomial ha de relacionar l'interval $(0,1)$ amb la recta real ja que el paràmetre de l'esperança μ pertany a aquest interval. Per fer-ho, és habitual utilitzar una de les tres funcions següents:

– logit:

$$\eta = \log \frac{\mu}{1-\mu}$$

Aquesta és la funció canònica de link i la que nosaltres utilitzarem quan apliquem aquesta distribució.

– probit:

$$\eta = \Phi^{-1}(\mu)$$

on, $\Phi(\cdot)$ és la funció de distribució de la Normal.

– complementari log-log:

$$\eta = \log - \log 1 - \mu$$

Com acabem de comprovar, la distribució Binomial pertany a la família exponencial i, per tant, podem aplicar la teoria que hem estudiat en aquest mateix capítol.

A.3 Models per a dades de comptatge

Les notacions del capítol 3 es poden consultar en aquesta secció de l'anex.

A.3.1 Sobredispersió

Per evitar la sobredispersió tenim diverses possibilitats:

- L'alternativa que podríem pensar seria considerar una **variància proporcional a l'esperança**:

  $$Var(Y) = \phi \mathbb{E}(Y) = \phi\mu$$

  Aquesta opció té l'inconvenient que quan apliquem el mètode dels $IWLS$ obtenim el mateix estimador dels pesos que quan es compleix la condició d'equidispersió, és a dir, que l'error estàndard es conservaria amb la presència de sobredispersió.

- Una altra opció seria considerar la metodologia de **Quasi-Poisson**. Aquesta metodologia té un coeficient més que estimar: el paràmetre de dispersió ϕ que s'estima a partir de les observacions mitjançant la formula (3.1), en comptes de prendre directament $\phi = 1$. Aquesta distribució utilitza els mateixos paràmetres que la Binomial Negativa, es diferencien en l'expressió de la variància: en el primer cas és una funció lineal respecte l'esperança i en el segon, una funció quadràtica.

- Tanmateix, l'opció més utilitzada és considerar una distribució **Binomial Negativa** degut a que en la seva funció de probabilitat intervenen dos paràmetres: l'esperença μ i un segon paràmetre θ que ens permetria ajustar la variància independentment de l'esperança.

Apèndix B

Apèndix de Base de dades

En aquesta secció adjuntarem totes les taules i gràfiques que s'han el·laborat per a portar a terme l'anàlisi exploratòria de dades, que per tal de no augmentar la llargària del document principal s'ha decidit que figurin a un apèndix.

B.1 Descripció de la base de dades

Antiguitat del carnet del conductor (X_2)

La taula de contingència següent mostra com es distribueixen el total de sinistres segons si el prenedor és una persona física o jurídica:

Tipus de prenedor	**Nº sinistres totals**			**Total**
	0	**1**	**2**	
Persones Físiques	28.240	1.599	97	29.936
Persones Jurídiques	57	6	1	64
Total	**28.297**	**1.605**	**98**	**30.000**

Taula B.1: Taula de contingència de les variables X_2 i Y

Temps d'exposició de la pòlissa (X_6)

A la Taula B.2 es mostra la taula de contingència de la variable any i la variable que indica els casos duplicats.

Identificació casos duplicats	**Any**			**Total**
	2014	**2015**	**2016**	
Casos primaris	1.525	2.251	23.327	27.103
Casos duplicats	26.575	36.281	10.941	73.797
Total	**28.100**	**38.532**	**34.268**	**100.900**

Taula B.2: Taula de contingència de les variables X_6 i la variable indicadora del tipus de cas

Tipus de vehicle (X_{10})

A continuació mostrem una taula de contingència del tipus de vehicle i la variable resposta, $Y =$ nombre total de sinistres.

Tipus de vehicle	**Nº sinistres totals**			**Total**
	0	**1**	**2**	
Turismes i derivats	28.297	1.605	98	30.000
Motocicletes	2.282	45	1	2.328
Furgonetes y Camions	1.182	91	11	1.284
Resta	650	6	0	656
Total	**32.411**	**1.747**	**110**	**34.268**

Taula B.3: Taula de contingència de les variables X_{10} i Y

B.2 Anàlisi exploratòria

B.2.1 Variables que caracteritzen al conductor habitual

Edat del conductor del vehicle i Antiguitat del carnet del conductor(X_1, X_2)

A la Figura B.1 mostrem l'scatter plot de les variables X_1 i X_2:

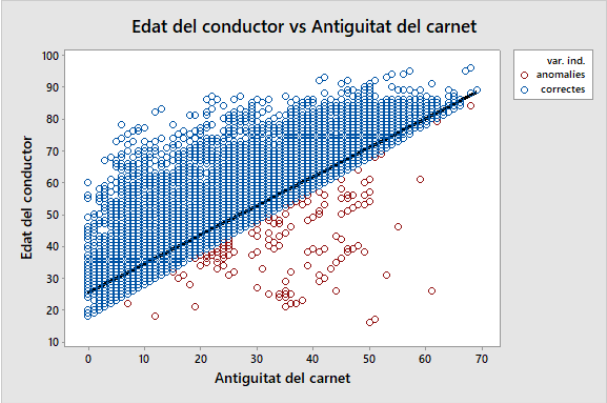

Figura B.1: Scatter plot de les variables X_1 i X_2 juntament amb la recta de regressió

A continuació, mostrarem l'scatter plot anterior un cop les dades anòmales han estat eliminades:

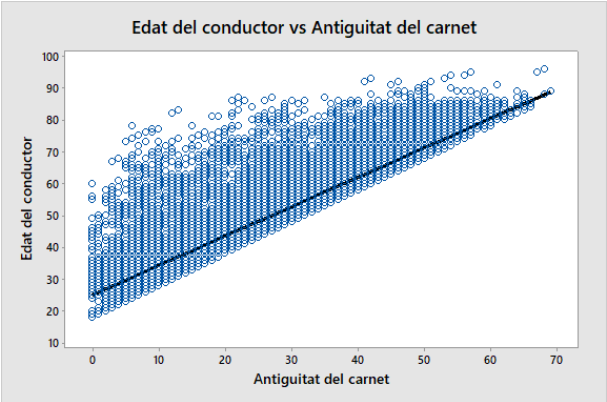

Figura B.2: Scatter plot de les variables X_1 i X_2 juntament amb la recta de regressió

La Figura B.3 mostra un gràfic combinat entre un diagrama de Pareto del número de pòlisses que té cada categoria i un diagrama de línies amb el seu percentatge de freqüència del nombre total de sinistres.

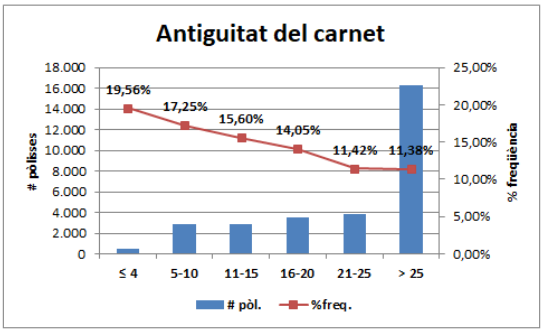

Figura B.3: Diagrama de Pareto associat a la variable X_2. EL percentatge correspon a la freqüència sinistral per cada categoria.

Coincidència de prenedor, conductor i propietari (X_3)

A la Figura B.4 s'ensenya una gràfica combinada del diagrama de Pareto que mostra el nombre total de pòlisses per cada categoria de la variable juntament amb la seva freqüència representada amb la línia poligonal.

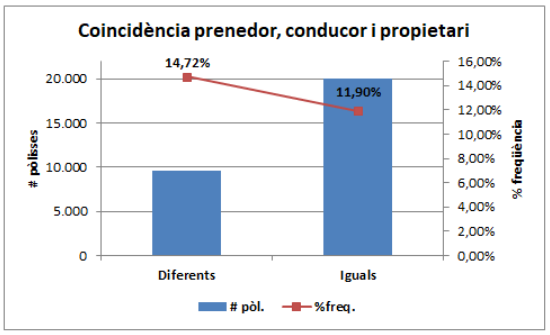

Figura B.4: Diagrama de Pareto de la variable X_3 juntament amb la poligonal que mostra la seva freqüència sinistral per categories.

Província de circulació(X_4)

A la Figura B.5 es presenta un diagrama de Pareto del nombre de pòlisses per cada categoria a l'eix principal (eix principal) i la línia poligonal amb el percentatge de la seva freqüència sinistral com a eix secundari (eix secundari).

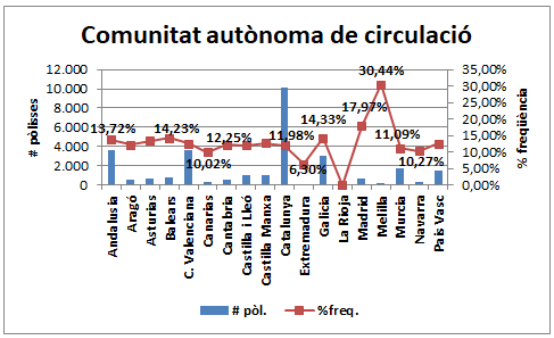

Figura B.5: Gràfica combinada d'un diagrama de Pareto que mostra el nombre de pòlisses per comunitat autònoma i la poligonal que presenta el seu percentatge de freqüència sinistral de la variable X_4

Temps d'exposició de la pòlissa(X_6)

A la Figura B.6 mostrem el gràfic combinat d'un diagrama de Pareto que representa el nombre de pòlisses de cada categoria i una poligonal amb la seva freqüència:

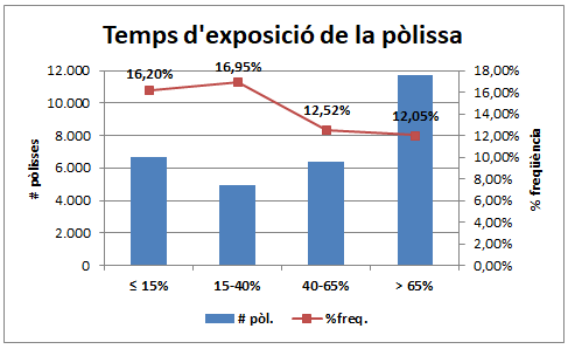

Figura B.6: Gràfic combinat de la variable X_6 d'un diagrama de Pareto que representa el nombre de pòlisses (eix principal) i la poligonal que mostra la seva freqüència (eix secundari)

Historial de sinistralitat del prenedor (X_7)

La Figura B.7 mostra un diagrama de Pareto que representa el nombre de pòlisses per les 5 categories de la variable i, per cadascuna, la poligonal descriu la seva freqüència sinistral:

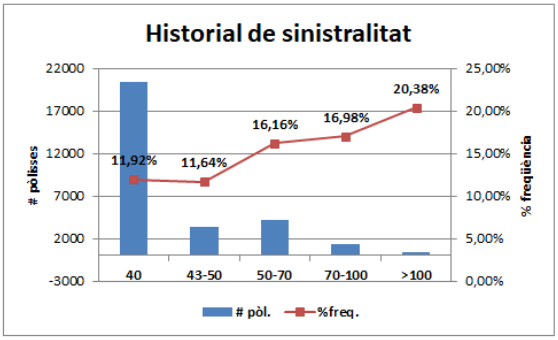

Figura B.7: Gràfica que combina un diagrama de Pareto amb el nombre de pòlisses per categoria i la poligonal que descriu la freqüència sinistral de la variable X_7

Nombre de pòlisses d'automòbils a ala companyia (X_8)

La Figura B.5 mostra un diagrama de Pareto del nombre de pòlisses per cada categoria a l'eix principal i la línia poligonal amb el percentatge de la seva freqüència sinistral com a eix secundari de la variable X_8

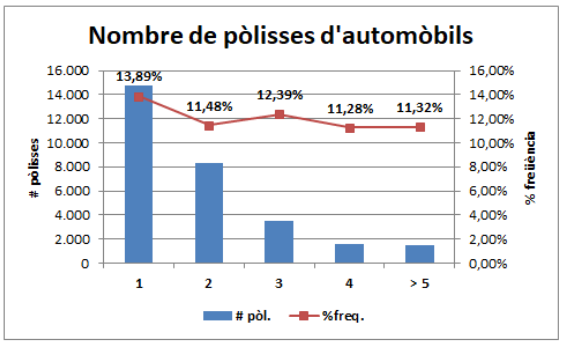

Figura B.8: Gràfica combinada d'un diagrama de Pareto que mostra el nombre de pòlisses per comunitat autònoma i la poligonal presenta el seu percentatge de freqüència sinistral de la variable X_8

Fraccionament del pagament (X_9)

A continuació, es mostra el gràfic de la variable X_9 que combina el diagrama de Pareto del nombre de pòlisses per cada categoria i la seva freqüència sinistral:

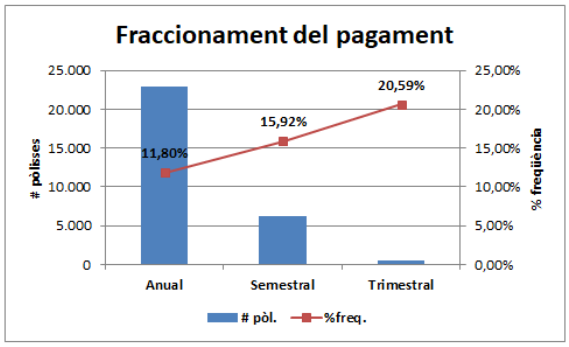

Figura B.9: Gràfic combinat del diagrama de Pareto del nombre de pòlisses i la poligonal de la freqüència de sinsitres total de la variable X_9

B.2.2 Variables que caracteritzen el vehicle assegurat

Tipus de vehicle (X_{10})

A la Figura B.10 figura el diagrama de Pareto que mostra el nombre total de pòlisses a l'eix principal i la poligonal que representa la freqüència sinistral de cada categoria a l'eix secundari:

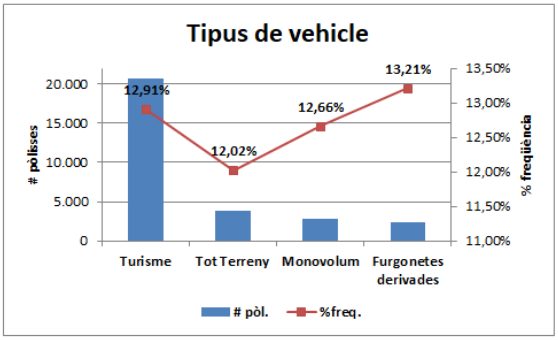

Figura B.10: Diagrama de Pareto que mostra el nombre de pòlisses per categoria a l'eix principal i la poligonal representa la freqüència sinistral a l'eix secundari

Combustible del vehicle (X_{11})

La Figura B.11 mostra una gràfica combinada amb el nombre de pòlisses que té cada categoria representat per les barres a l'eix principal i la seva freqüència en la línia poligonal com a eix secundari:

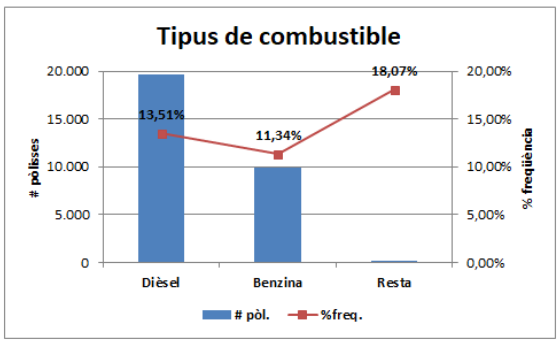

Figura B.11: Gràfica que combina un diagrama de Pareto del nombre de pòlisses que té la variable X_9 per cada categoria amb la seva freqüència sinistral a l'eix principal

Antiguitat del vehicle (X_{12})

A la Figura B.12, les barres representant el total de pòlisses per cada categoria i la poligonal el percentatge de freqüència sinistral que tenen:

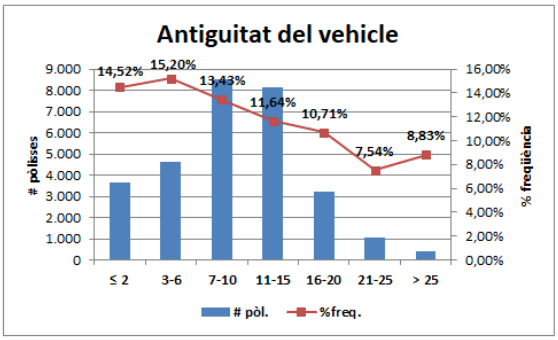

Figura B.12: Gràfica combinada de l'diagrama de Pareto del nombre de pòlisses per cada categoria a l'eix principal i el seu percentatge de freqüència a l'eix de la dreta

La Figura B.13 conté l'scatter plot de les variables X_2 i X_9 per saber quina relació hi ha entre elles:

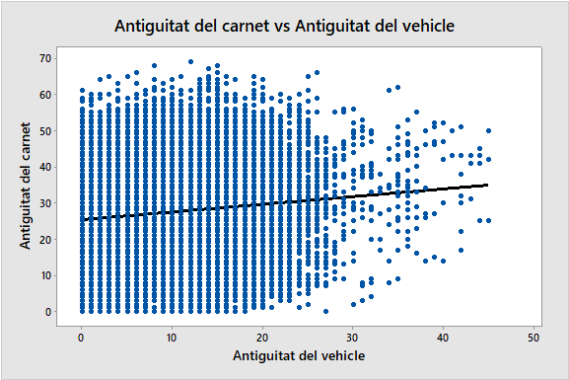

Figura B.13: Diagrama de dispersió de les variables X_2 i X_9

Valor total del vehicle (X_{13})

A la Figura B.14 tenim el gràfic de la variable X_{10} que conté el diagrama de Pareto amb el nombre de pòlisses de cada categoria en l'eix principal i la poligonal que representa la seva freqüència sinistral:

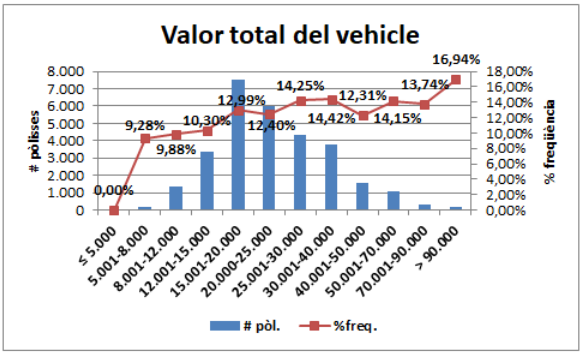

Figura B.14: Diagrama de Pareto del nombre de pòlisses per cada categoria a l'eix secundari combinat amb el seu percentatge de freqüència sinistral a l'eix principal

Valor total del vehicle, Potència i Cilindrada (X_{13}, X_{14} i X_{16})

La Figura B.15 conté l'scatter plot d'aquestes dues variables i el valor total del vehicle per comprovar, a més, si té sentit considerar que com més car és el valor del vehicle més potencia té :

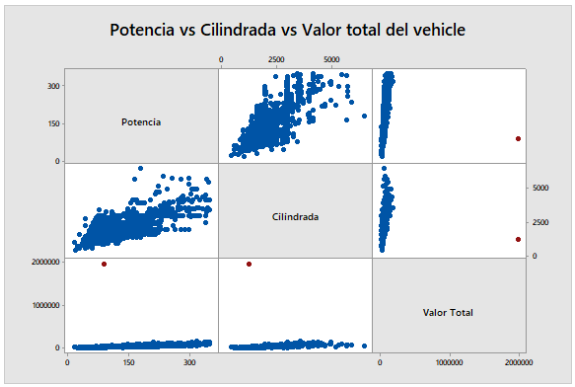

Figura B.15: Matriu de dispersió de les variables X_{12}, X_{14} i X_{11}

A continuació, mostrarem com queda l'scatter plot quan eliminem el vehicle de valor $1.981.494$ ja que és un outlier i fa augmentar la variabiitat entre les observacions:

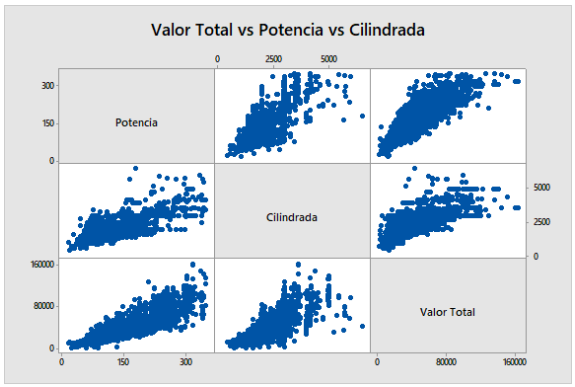

Figura B.16: Matriu de dispersió de les variables X_{12}, X_{14} i X_{11}

Quocient pes/potencia (X_{17})

A la Figura B.17 mostrem com evoluciona la freqüència sinistral en les 12 categories d'aquesta variable en l'eix secundari i el nombre de pòlisses que té cadascuna d'elles en l'eix principal:

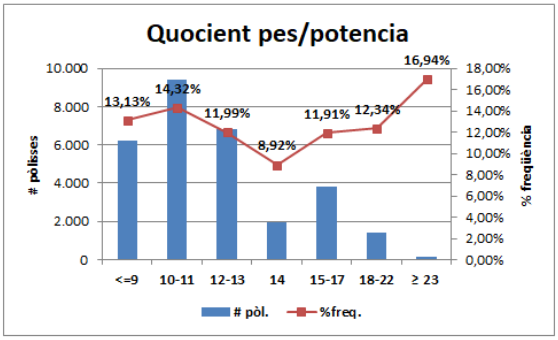

Figura B.17: Gràfica combinada d'un diagrama de Pareto amb el nombre de pòlisses de cada categoria a l'eix secundari i la seva freqüència a l'eix principal de la variable X_{17}

Apèndix C

Apèndix de la Modelització

C.1 Model Lineal amb resposta transformada

Com hem esmentat a la introducció, aquest treball ens ha donat l'oportunitat de conèixer millor el software estadístic R. A la assignatura d'Estadística vam aprendre el seu ús molt per sobre i el vam aplicar en algunes pràctiques. En aquest apèndix mostrarem tots els procediments i resultats necessaris per realitzar el capítol 5.
Per poder treballar la base de dades, el primer pas serà importa l'arxiu .xlsx i després seleccionar només aquelles variables que vam esmentar que treballaríem finalment:

```
# Importem la base de dades
library(readxl)
dades <- read_excel("C:/Users/NuriiiBG/Desktop/TFG_definitivo/DEFINITIVO/dades.xlsx")
View(dades)

# Denotem les columnes com a variables
attach(dades)

# Seleccionem només aquelles variables amb les que anem a treballar
data <- dades[, c(50, 43, 4, 5, 13, 16, 39,20:22, 25, 28, 30, 34, 40)]
View(data)
```

A continuació mostrarem els resultats per als models lineals, utilitzant com a variable resposta els canvis esmentats al capítol 5.
Recordem que el canvi 1 és $y_1 = \sqrt{y + \frac{3}{8}}$:

```
##############
#MODELITZACIÓ#
##############

#################
#0.NORMALITZACIÓ#
#################

# Primer canvi: y1 = sqrt(y+3/8)

ml1 <- lm(sqrt(sintot+3/8)~ tempexp + edacar + as.factor(mismafig) + as.factor(procir) + hpsrct
          + autpol + antveh + valtot + as.factor(pagpol),
          data = data)
summary(ml1)
```

i els resultats obtinguts:

```
##
## Call:
## lm(formula = sqrt(sintot + 3/8) ~ tempexp + edacar + as.factor(mismafig) +
##     as.factor(procir) + hpsrct + autpol + antveh + valtot + as.factor(pagpol),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12264 -0.04610 -0.03335 -0.01306  0.91288
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.203e-01  5.881e-03 105.473  < 2e-16 ***
## tempexp                 6.086e-02  2.919e-03  20.848  < 2e-16 ***
## edacar                 -2.513e-03  5.577e-04  -4.506 6.64e-06 ***
## as.factor(mismafig)1 -5.621e-03  1.758e-03  -3.197  0.00139 **
## as.factor(procir)2     5.120e-03  2.616e-03   1.957  0.05039 .
## as.factor(procir)3     3.859e-03  5.878e-03   0.657  0.51149
## as.factor(procir)4     4.283e-03  5.402e-03   0.793  0.42786
## as.factor(procir)5     7.471e-03  5.096e-03   1.466  0.14264
## as.factor(procir)6     2.904e-03  2.628e-03   1.105  0.26913
## as.factor(procir)7    -3.609e-03  7.144e-03  -0.505  0.61342
## as.factor(procir)8     2.047e-03  6.093e-03   0.336  0.73686
## as.factor(procir)9     3.117e-03  4.376e-03   0.712  0.47631
## as.factor(procir)10    5.282e-03  4.549e-03   1.161  0.24557
## as.factor(procir)11   -1.318e-02  1.155e-02  -1.141  0.25370
## as.factor(procir)12    7.552e-03  2.785e-03   2.712  0.00669 **
## as.factor(procir)13   -3.330e-02  4.065e-02  -0.819  0.41260
## as.factor(procir)14    1.110e-02  5.330e-03   2.082  0.03735 *
## as.factor(procir)15    4.424e-02  9.556e-03   4.630 3.68e-06 ***
## as.factor(procir)16    2.139e-05  3.537e-03   0.006  0.99518
## as.factor(procir)17   -2.174e-03  8.084e-03  -0.269  0.78800
## as.factor(procir)18    2.325e-03  3.804e-03   0.611  0.54113
## hpsrct                 2.631e-04  6.614e-05   3.978 6.97e-05 ***
## autpol                -2.143e-03  5.404e-04  -3.965 7.36e-05 ***
## antveh                -2.610e-03  5.914e-04  -4.414 1.02e-05 ***
## valtot                 1.307e-03  4.386e-04   2.980  0.00289 **
## as.factor(pagpol)2     8.548e-03  1.956e-03   4.371 1.24e-05 ***
## as.factor(pagpol)3     1.664e-02  6.262e-03   2.657  0.00789 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1347 on 29705 degrees of freedom
## Multiple R-squared:  0.02034,    Adjusted R-squared:  0.01948
## F-statistic: 23.72 on 26 and 29705 DF,  p-value: < 2.2e-16
```

El canvi 2 és $y_2 = y^{2/3}$:

```
# Segon canvi: y2 = y^(2/3)

ml2 <- lm(sintot^(2/3)~ tempexp + edacar + as.factor(mismafig) + as.factor(procir) + hpsrct
          + autpol + antveh + valtot + as.factor(pagpol), data = data)
summary(ml2)
```

i el resultat:

```
##
## Call:
## lm(formula = sintot^(2/3) ~ tempexp + edacar + as.factor(mismafig) +
##     as.factor(procir) + hpsrct + autpol + antveh + valtot + as.factor(pagpol),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21677 -0.08194 -0.05934 -0.02328  1.55913
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.0141560  0.0104341   1.357  0.17488
## tempexp               0.1081105  0.0051792  20.874  < 2e-16 ***
## edacar               -0.0044369  0.0009894  -4.485 7.33e-06 ***
## as.factor(mismafig)1 -0.0099737  0.0031188  -3.198  0.00139 **
## as.factor(procir)2    0.0090724  0.0046418   1.954  0.05065 .
## as.factor(procir)3    0.0070716  0.0104283   0.678  0.49770
## as.factor(procir)4    0.0077735  0.0095834   0.811  0.41729
## as.factor(procir)5    0.0134410  0.0090402   1.487  0.13708
## as.factor(procir)6    0.0051804  0.0046619   1.111  0.26648
## as.factor(procir)7   -0.0064427  0.0126732  -0.508  0.61120
## as.factor(procir)8    0.0035962  0.0108094   0.333  0.73937
## as.factor(procir)9    0.0055600  0.0077631   0.716  0.47387
## as.factor(procir)10   0.0095581  0.0080695   1.184  0.23624
## as.factor(procir)11  -0.0233378  0.0204885  -1.139  0.25469
## as.factor(procir)12   0.0133324  0.0049401   2.699  0.00696 **
## as.factor(procir)13  -0.0592009  0.0721119  -0.821  0.41168
## as.factor(procir)14   0.0195283  0.0094560   2.065  0.03892 *
## as.factor(procir)15   0.0771002  0.0169533   4.548 5.44e-06 ***
## as.factor(procir)16   0.0001041  0.0062746   0.017  0.98677
## as.factor(procir)17  -0.0039212  0.0143418  -0.273  0.78454
## as.factor(procir)18   0.0042352  0.0067485   0.628  0.53029
## hpsrct                0.0004657  0.0001173   3.968 7.25e-05 ***
## autpol               -0.0037858  0.0009587  -3.949 7.87e-05 ***
## antveh               -0.0046521  0.0010492  -4.434 9.29e-06 ***
## valtot                0.0023155  0.0007782   2.976  0.00293 **
## as.factor(pagpol)2    0.0152711  0.0034695   4.402 1.08e-05 ***
## as.factor(pagpol)3    0.0300200  0.0111094   2.702  0.00689 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.239 on 29705 degrees of freedom
## Multiple R-squared:  0.02036,    Adjusted R-squared:  0.0195
## F-statistic: 23.74 on 26 and 29705 DF,  p-value: < 2.2e-16
```

C.2 Regressió de Poisson:

Apliquem l'eina dels MLG a la distribució de Poisson de la variable resposta:

```
# Regressió de Poisson
pois <- glm(sintot ~ offset(log(tempexp)) + edacar + as.factor(mismafig) + as.factor(procir)
            + hpsrct + autpol + antveh + valtot + as.factor(pagpol), data = data,
            family = "poisson")
summary(pois)
```

els resultats obtinguts són:

```
##
## Call:
## glm(formula = sintot ~ offset(log(tempexp)) + edacar + as.factor(mismafig) +
##     as.factor(procir) + hpsrct + autpol + antveh + valtot + as.factor(pagpol),
##     family = "poisson", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9779  -0.4077  -0.3322  -0.1798   3.6472
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.931060   0.162255 -11.901  < 2e-16 ***
## edacar                 -0.070936   0.015845  -4.477 7.58e-06 ***
## as.factor(mismafig)1   -0.180843   0.051788  -3.492 0.000479 ***
## as.factor(procir)2      0.149581   0.077851   1.921 0.054684 .
## as.factor(procir)3      0.077484   0.184683   0.420 0.674811
## as.factor(procir)4      0.106932   0.158372   0.675 0.499550
## as.factor(procir)5      0.209406   0.147895   1.416 0.156801
## as.factor(procir)6      0.089327   0.081335   1.098 0.272090
## as.factor(procir)7     -0.136870   0.246540  -0.555 0.578783
## as.factor(procir)8      0.050238   0.184542   0.272 0.785445
## as.factor(procir)9      0.091085   0.136222   0.669 0.503716
## as.factor(procir)10     0.138737   0.138297   1.003 0.315773
## as.factor(procir)11    -0.615513   0.501855  -1.226 0.220020
## as.factor(procir)12     0.223276   0.081007   2.756 0.005847 **
## as.factor(procir)13   -10.321204 129.208404  -0.080 0.936332
## as.factor(procir)14     0.309529   0.139668   2.216 0.026679 *
## as.factor(procir)15     0.949077   0.190866   4.972 6.61e-07 ***
## as.factor(procir)16    -0.014701   0.115471  -0.127 0.898692
## as.factor(procir)17    -0.077054   0.270770  -0.285 0.775971
## as.factor(procir)18     0.064980   0.116345   0.559 0.576495
## hpsrct                  0.006914   0.001704   4.058 4.95e-05 ***
## autpol                 -0.077101   0.018720  -4.119 3.81e-05 ***
## antveh                 -0.083943   0.018351  -4.574 4.78e-06 ***
## valtot                  0.042165   0.013170   3.202 0.001367 **
## as.factor(pagpol)2      0.245254   0.055600   4.411 1.03e-05 ***
## as.factor(pagpol)3      0.354987   0.153447   2.313 0.020700 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 9778.0  on 29731  degrees of freedom
## Residual deviance: 9588.9  on 29706  degrees of freedom
## AIC: 13074
##
## Number of Fisher Scoring iterations: 11
```

C.3 Model de Poisson zero inflat:

Degut a l'excés de zeros provem si la distribució discreta Poisson zero inflada fita millor les nostres observacions:

```
# Model de Poisson zero inflat
library(pscl)
zip <- zeroinfl(sintot ~ offset(log(tempexp)) + edacar + as.factor(mismafig) + as.factor(procir)
           + hpsrct + autpol + antveh + valtot + as.factor(pagpol)|1, data = data,
           dist = "poisson")
summary(zip)
```

i el resultat:

```
##
## Call:
## zeroinfl(formula = sintot ~ offset(log(tempexp)) + edacar + as.factor(mismafig) +
##     as.factor(procir) + hpsrct + autpol + antveh + valtot + as.factor(pagpol) |
##     1, data = data, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.6411 -0.2839 -0.2324 -0.1269 30.3503
##
## Count model coefficients (poisson with log link):
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.611983   0.191127  -8.434  < 2e-16 ***
## edacar                -0.071139   0.016145  -4.406 1.05e-05 ***
## as.factor(mismafig)1  -0.180641   0.052707  -3.427  0.00061 ***
## as.factor(procir)2     0.148826   0.079206   1.879  0.06025 .
## as.factor(procir)3     0.075309   0.187639   0.401  0.68816
## as.factor(procir)4     0.105784   0.161104   0.657  0.51143
## as.factor(procir)5     0.210650   0.150549   1.399  0.16175
## as.factor(procir)6     0.089669   0.082638   1.085  0.27788
## as.factor(procir)7    -0.137234   0.249753  -0.549  0.58268
## as.factor(procir)8     0.050177   0.187557   0.268  0.78906
## as.factor(procir)9     0.089586   0.138344   0.648  0.51727
## as.factor(procir)10    0.139089   0.140587   0.989  0.32249
## as.factor(procir)11   -0.613811   0.506263  -1.212  0.22535
## as.factor(procir)12    0.224825   0.082468   2.726  0.00641 **
## as.factor(procir)13  -10.321941 213.009137  -0.048  0.96135
## as.factor(procir)14    0.308409   0.142645   2.162  0.03061 *
## as.factor(procir)15    0.969651   0.198934   4.874 1.09e-06 ***
## as.factor(procir)16   -0.015765   0.117124  -0.135  0.89293
## as.factor(procir)17   -0.082418   0.274302  -0.300  0.76382
## as.factor(procir)18    0.066961   0.118233   0.566  0.57116
## hpsrct                 0.007030   0.001765   3.982 6.84e-05 ***
## autpol                -0.076964   0.018941  -4.063 4.84e-05 ***
## antveh                -0.084015   0.018652  -4.504 6.66e-06 ***
## valtot                 0.042423   0.013418   3.162  0.00157 **
## as.factor(pagpol)2     0.246558   0.056645   4.353 1.34e-05 ***
## as.factor(pagpol)3     0.349653   0.157258   2.223  0.02619 *
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.9588     0.3476  -2.758  0.00581 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 41
## Log-likelihood: -6506 on 27 Df
```

C.4 Comparació dels models

Hem fet una taula comparativa d'ambdós models. A continuació s'exposa com es calcula cada terme:

```r
# Taula comparativa

# Poisson
lpois <- logLik(pois)
aic_pois <- AIC(pois)
dev_pois <- deviance(pois)
pear_pois <- sum(residuals(pois, type = "pearson")^{2})/(29732-9)

# ZIP
lzip <- logLik(zip)
aic_zip <- AIC(zip)
dev_zip <- deviance(zip)
pear_zip <- sum(residuals(zip, type = "pearson")^{2})/(29732-10)
```

Per acabar, només queda mostrar la gràfica dels residus pels dos models:

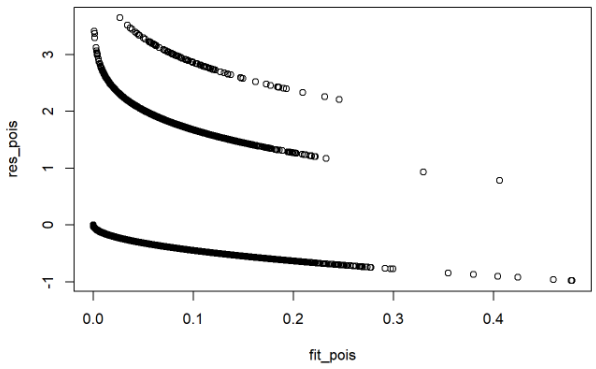

Figura C.1: Gràfica dels residus de la regressió de Poisson

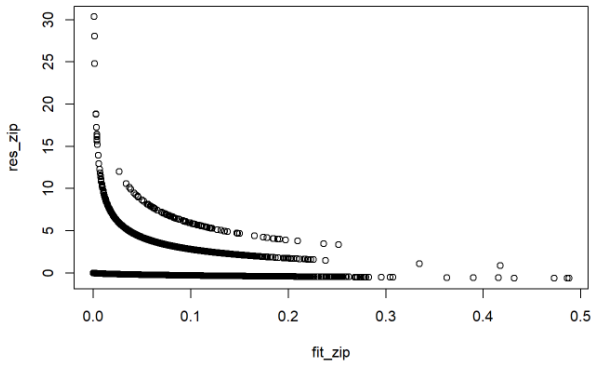

Figura C.2: Gràfica dels residus pel models de Poisson zero inflat

Bibliografia

[1] RAWLINGS, JOHN O., PANTULA, SASTRY G., DICKEY, DAVID A..*Review of Simple Regression*. En:CASELLA, GEORGE, FIENBERG, STEPHEN, OLKIN, INGRAM, editors.*Applied Regression Analysis: A Research Tool*.Second Edition. New York, USA: Springer-Verlag;1998.p.1-30.

[2] RAWLINGS, JOHN O., PANTULA, SASTRY G., DICKEY, DAVID A..*Multiple Regression in matrix notation*. En:CASELLA, GEORGE, FIENBERG, STEPHEN, OLKIN, INGRAM, editors.*Applied Regression Analysis: A Research Tool*.Second Edition. New York, USA: Springer-Verlag;1998.p.1-30.

[3] FARAWAY, JULIAN J.. *Estimation*. En:CHATFIELD, CHRIS, TANNER, MARTIN, ZIDEK, JIM, editors. *Linear Models with R*. Boca Raton, London, NY, Washington, D.C.: Chapman & Hall/CRC; 2005. p.12-23.

[4] FARAWAY, JULIAN J.. *Inference*. En:CHATFIELD, CHRIS, TANNER, MARTIN, ZIDEK, JIM, editors. *Linear Models with R*. Boca Raton, London, NY, Washington, D.C.: Chapman & Hall/CRC; 2005. p.18-53.

[5] CARMONA, FRANCESC. *Las condiciones*. En: *Modelos Lineales*. Universitat de Barcelona 2001. p.7-14. Disponible en: http://www.ub.edu/stat/docencia/Diplomatura/ModelsLineals/regre.pdf

[6] CARMONA, FRANCESC. *Estimación*. En: *Modelos Lineales*. Universitat de Barcelona 2001. p.18-28. Disponible en: http://www.ub.edu/stat/docencia/Diplomatura/ModelsLineals/regre.pdf

[7] CARMONA, FRANCESC. *Funciones paramètricas estimables*. En: *Modelos Lineales*. Universitat de Barcelona 2001. p.33-56. Disponible en: http://www.ub.edu/stat/docencia/Diplomatura/-ModelsLineals/regre.pdf

[8] CARMONA, FRANCESC. *Regresión*. En: *Modelos Lineales*. Universitat de Barcelona, 2001. p.59-70. Disponible en: http://www.ub.edu/stat/docencia/Diplomatura/ModelsLineals/regre.pdf

[9] PÉREZ CASSANY, MARTA. *Linear and Generalized Linear Models slides*, [Model Lineal i Lineal Generalitzats], Màster MESIO, Universitat Politècnica de Catalunya. 2017.

[10] RODRÍGUEZ, GERMÁN. *Appendix B: Generalized Linear Model Theory*. En: *Lecture Notes on Generalized Linear Models*. Princeton University, 2007. Disponible en: http://data.princeton.edu/wws509/notes/a2.pdf

[11] DOBSON, ANNETTE J. *Model Fitting*. En: CHATFIELD, C., ZIDEK, J. editors. *An introduction to Generalized Linear Models*. Second Edition. Boca Raton, London, NY, Washington, D.C.: Chapman & Hall/CRC; 2002. p.25-49.

[12] DOBSON, ANNETTE J. *Exponential Family and Generalized Linear Models*. En: CHATFIELD, C., ZIDEK, J. editors. *An introduction to Generalized Linear Models*. Second Edition. oca Raton, London, NY, Washington, D.C.: Chapman & Hall/CRC; 2002. p.50-62.

[13] DOBSON, ANNETTE J. *Estimation*. En: CHATFIELD, C., ZIDEK, J. editors. *An introduction to Generalized Linear Models*. Second Edition. Boca Raton, London, NY, Washington, D.C.: Chapman & Hall/CRC; 2002. p.63-74.

[14] DOBSON, ANNETTE J. *Count Data, Poisson Regression and Log-Linear Models*. En: CHATFIELD, C., ZIDEK, J. editors. *An introduction to Generalized Linear Models*. Second Edition. Boca Raton, London, NY, Washington, D.C.: Chapman & Hall/CRC; 2002. p.156-175.

[15] MCCULLAGH, P., NELDER, J.A.*An outline of generalized linear models*. En: MCCULLAGH, P., NELDER, J.A. editors. *Generalized Linear Models*. Second Edition. London, NY: Chapman & Hall; 1983. p.21-43.

[16] MCCULLAGH, P., NELDER, J.A. *Log-linear models*. En: MCCULLAGH, P., NELDER, J.A. editors. *Generalized Linear Models*. Second Edition. London, NY: Chapman & Hall; 1983. p.193-235.

[17] MARÍN, JUAN MIGUEL. *Apuntes Tema 3: Modelos lineales generalizados*. Universidad Carlos III, Madrid. Disponible en: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Categor/-Tema3Cate.pdf

[18] PÉREZ CASSANY, MARTA. *GLM: Binomial Response slides*,[Model Lineal i Lineal Generalitzats], Màster MESIO, Universitat Politècnica de Catalunya. 2017.

[19] PÉREZ CASSANY, MARTA. *GLM: Poisson Response slides*,[Model Lineal i Lineal Generalitzats], Màster MESIO, Universitat Politècnica de Catalunya. 2017.

[20] RODRÍGUEZ, GERMÁN. *Models for Count data with Overdisperion*. En: *Lecture Notes on Generalized Linear Models*. Princeton University, 2007. Disponible en: http://data.princeton.edu/wws509/notes/c4a.pdf

[21] COLIN, A. CAMERON, TRIVEDI, PRAVIN K. *Models of Count Data*. En: COLIN, A. CAMERON, TRIVEDI, PRAVIN K. editors. MICROECONOMETRICS: METHODS AND APPLICATIONS. New York: Cambridge University. 2005. p. 665-674.

[22] ZEILEIS, ACHIM, KLEIBER, CHISTIAN, JACKMAN, SIMON *Regression Models for Count Data in R* Journal of Statisttical Software. 2008; 27(8): 25. Disponible en: https://www.jstatsoft.org/article/view/v027i08

[23] GILENKO, EVGENII V., MIRONOVA, ELENA A. *Modern claim frequency and claim severity models: An application to the Russian motor own damage insurance market*. Cogent Economic & Finance. 2017; 5(1311097). Disponible en: http://www.tandfonline.com/doi/full/10.1080/23322039.2017.1311097